DOCUMENT RESUME

ED 135 801 TM 005 718

AUTHOR Cronbach, Lee J.; And Others
TITLE Research on Classrooms and Schools: Formulation of Questions, Design and Analysis.
INSTITUTION Stanford Univ., Calif. Stanford Evaluation Consortium.
SPONS AGENCY Russell Sage Foundation, New York, N.Y.; Spencer Foundation, Chicago, Ill.
PUB DATE Jul 76
NOTE 243p.
AVAILABLE FROM Stanford Evaluation Consortium, School of Education, Stanford University, Stanford, California 94305 ($1.00)

EDRS PRICE MF-$0.83 HC-$12.71 Plus Postage.
DESCRIPTORS Analysis of Covariance; Classroom Research; *Classrooms; *Educational Research; Mathematical Models; *Research Design; Research Methodology; *Research Problems; Sampling; *Schools; Social Science Research; Statistical Analysis
IDENTIFIERS Aggregation Effects; *Aptitude Treatment Interaction

ABSTRACT

Alternative ways of analyzing data from Aptitude Treatment Interactions were examined over a two-year period. In light of past arguments the author maintains that the questions surrounding aggregation have been badly posed and that the customary methods of analysis were either incorrect or subject to misinterpretation. Therefore, the majority of studies of educational effects--whether classroom experiments, or evaluations of programs or surveys--have collected and analyzed data in ways that conceal more than they reveal. The established methods have generated false conclusions in many studies. Further, the traditional research strategy which pits substantive hypotheses against a null hypothesis and requires statistical significance of effects can rarely be used in educational research. Samples large enough to detect strong, but probabilistic effects are likely to be prohibitively costly. (Author/MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED135801

TM

OCCASIONAL PAPERS OF THE STANFORD
Evaluation Consortium

Stanford University, Stanford, California, 94305

TM005 718

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Stanford
Evaluation Consortium
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

RESEARCH ON CLASSROOMS AND SCHOOLS:
FORMULATION OF QUESTIONS, DESIGN, AND ANALYSIS

Lee J. Cronbach

with the assistance of

Joseph E. Deken and Noreen Webb

July, 1976

The Stanford Evaluation Consortium is a group of faculty members and students concerned with the improvement of evaluation of educational and social-service programs. The Occasional Papers represent the views of the authors as individuals. Comments and suggestions for revision are invited. The papers should not be quoted or cited without the specific permission of the author; they are automatically superseded upon formal publication of the material.

Additional copies of this paper are available for $1.00 each from the Stanford Evaluation Consortium, School of Education, Stanford University, Stanford, California 94305.

Table of Contents

Preface 1

1. Introduction to the problem 1.1
    - From a statistical issue to a substantive issue 1.1
        - The Abt Follow Through report 1.3a
        - The need to disentangle effects 1.9
    - Psychological bias and sociological bias 1.14
        - A debate in educational psychology 1.15
        - Debates within sociology 1.16
    - Units of analysis? of treatment? of theory? 1.18
        - Units within hierarchies 1.22
        - Units in areal analysis 1.23

2. Units in various research contexts 2.1
    - The problem as seen in research on ATI 2.1
        - Sample size for regression analysis 2.2
        - The Maier-Jacobs study 2.3
        - Harvard Project Physics 2.5
        - Three kinds of process 2.6
    - Aggregation effects 2.11
    - Ecological psychology 2.16
    - Evaluative studies and school-effect studies 2.17
    - Extrapolation in interpretation 2.23

3. A mathematical model 3.1
    - Definition of components 3.1
        - Interpretation of components 3.4
    - Partitioning variance 3.8

Choices made in forming the model 3.9
  Direction of decomposition 3.9
  Nonlinearity 3.10
Effects of aggregating data 3.11
  Special case with linear assumption 3.12
  Meaning of ϕ 3.15
  Comparing β_b and β_w 3.16
  Implication for ATI research 3.21
  Aggregation effects with multiple discriminants 3.22

4. The reference population and its parameters 4.1
Alternative models for statistical inference 4.1
  Collectives distinct, persons fixed 4.2
  Collectives nested within local populations 4.3
  The independence assumption 4.4
  Choice among models 4.5
Weights that define parameters 4.7
Illustrative statistics for populations of collectives 4.11
  Head Start 4.11
  School districts in California 4.12
A problem of estimation 4.13

5. Illustrative ATI studies 5.1
The Anderson study 5.1
  A weighting decision 5.2
  Regressions of ZACH on PRECOM 5.3
  Regressions of ZACH on ABIL 5.8
Cooperative Reading data 5.11
  Plan of the studies 5.11

- The original analysis across projects 5.12
- Half-class as unit of analysis? 5.13
- The original analysis within projects 5.14
- Procedures in our reanalysis 5.16
- Results of our analysis 5.18
- Conclusions regarding units of analysis 5.22

Head Start Planned Variation 5.23

6. Disattenuating regression slopes 6.1

  Within- and between-groups reliabilities 6.2
  - Case I 6.3
  - Case II 6.5
  - Case III 6.5

  General remarks 6.8

7. Statistical inference 7.1

  Sampling error of a mean 7.2

  The between-groups regression 7.3

  The within-groups regression 7.4

8. Analysis of covariance 8.1

  Design 1. Collectives nested in treatments 8.2
  - Alternative adjustments 8.3
  - Cooperative Reading data 8.8
  - Follow Through data 8.9

  Design 2. Treatments crossed with blocks; collectives nested 8.10
  - The plan of the Head Start study 8.11
  - Alternative analyses, assuming homogeneity of regression 8.13
  - Alternative analyses, recognizing heterogeneity of regressions 8.17

9. Multivariate considerations 9.1

- Simple correlations 9.1
  - Correlations of reading outcomes 9.2
- Component analysis and factor analysis 9.6
  - Analysis of correlations 9.11
  - Factoring covariances 9.13
  - Slatin's analyses 9.15
  - Suggestions 9.16
  - Empirical test construction 9.19
- Multiple regression and related techniques 9.20
  - A school-effects model 9.20
  - Partialling 9.22

10. The Road Ahead 10.1

Preface

If any fraction of the argument herein is correct, educational research -- and a great deal of social science -- is in serious trouble. The implications of my analysis can be put bluntly:

1. The majority of studies of educational effects -- whether classroom experiments, or evaluations of programs, or surveys -- have collected and analyzed data in ways that conceal more than they reveal. The established methods have generated false conclusions in many studies.

2. The traditional research strategy -- pitting substantive hypotheses against a null hypothesis and requiring statistical significance of effects -- can rarely be used in educational research. Samples large enough to detect strong but probabilistic effects are likely to be prohibitively costly.

This work began with a perplexity regarding an aspect of an isolated kind of research on instruction. Funds from the Spencer Foundation enabled me to examine with two assistants, over a two-year period, some alternative ways of analyzing data from studies of Aptitude × Treatment interactions. When the time came to write a final report, I discovered that the problem was larger, more important, and less tractable than we had assumed when we were concentrating on ATI studies and techniques (Cronbach & Snow, 1976).

Sociologists have discussed and debated for many years about the legitimacy of explaining data in terms of "context effects" or "compositional effects". Similar but less extensive discussion is to be found in the literature of political science, and comparable questions arise in trying to reconcile microeconomics with macroeconomics. That literature has been virtually ignored by educational research workers and by psychologists, except in the "school-effect" studies of the past ten years. If this monograph does no more than alert my colleagues to the perils of ignoring issues of aggregation, that would be sufficient justification.

As I studied the past arguments, I came to think that the questions surrounding aggregation have been badly posed, and that the customary methods of analysis were either incorrect or subject to misinterpretation. Hence I am addressing a broad audience of social scientists rather than merely those in educational research.

As I began to set down my ideas and puzzlements, the plot thickened with each passing week. Many recent publications and unpublished reports came to my attention; I particularly credit a paper by Walt Haney (1974b) for its seminal influence. I tried parts of the argument on knowledgeable colleagues, and those interchanges moved my thinking further.

It will be evident from the physical form of the paper that it is a draft still undergoing revision.

I have decided to distribute it in this form because of my conviction that these issues are of vital importance and that it would be counterproductive to delay the discussion until my argument is polished. Millions of dollars are going into evaluation studies each year; it would be a sufficient short-run contribution to persuade sponsors and investigators to think hard about the questions raised here. I have not resolved the problems presented by aggregate phenomena; it is my intention here to stir up debate and to encourage proposals from others. I invite -- nay, beseech -- comments and counterarguments from those who receive this paper.

This project originated out of a concern with Aptitude × Treatment interactions. The procedure in ATI research is to calculate an outcome-on-aptitude regression for some teaching method, with the intent of discovering and explaining differences in regression slopes for alternative methods. During the same period I became involved in theoretical aspects of analysis of covariance. In that analysis, regression slopes have been regarded as instrumental rather than as of primary interest. As the Abt example in Section 1 shows, the issue of units of analysis arises there also. Thirdly, Leigh Burstein has completed a doctoral dissertation on the aggregation problem, as seen particularly in studying regressions calculated in educational sociology. I have served as chairman of his dissertation committee and find that experience influencing my thinking here.

My two assistants, Joseph Deken and Noreen Webb, played an important role in developing materials for this monograph. Miss Webb took primary responsibility for the illustrative data analyses and the data-processing methods, and Mr. Deken led the way in the statistical theorizing. Neither of them is to be held responsible for the present content of the paper. Analyses of California Assessment data were made by David Rogosa, under support from the State, and Lynne Gray assisted in analyzing the Featherstone data. David E. Wiley was good enough to work through the entire manuscript and did much to correct and extend my thoughts. Conversations with Leigh Burstein, Merrill Carlsmith, Dan Davis, Mike Hannan, and David Rogosa were helpful; likewise, I thank Dudley Duncan, Robert Hauser, and Herbert Walberg for suggestions in correspondence.

The revision of the report prepared for the Spencer Foundation and its reproduction and distribution was supported by the Stanford Evaluation Consortium under a grant from the Russell Sage Foundation.

1. Introduction to the Problem

From a statistical issue to a substantive issue

If a contour map of the region from Maine to Maryland were prepared with a six-inch contour interval, a person inspecting the map would not perceive the Atlantic Ocean. Some such remark was the lead sentence of an article in Science some years back. I have been unable to locate the article, and the contour mentioned may have been six feet, not six inches. However the writer phrased it, the point is that a fine-grain analysis can overlook large configurations in the data. It is equally true that too gross an analysis can conceal important relationships. Nor are all sins those of omission; investigation on the wrong scale can positively distort relations.

It is conventional in psychology and biology to regard the single organism as the object of investigation, and educational research workers took over that point of view. They frame hypotheses in terms of individuals and base their analysis on individual scores, though they do not always make the person the sampling unit. Only subsequent to the Coleman report of 1966 did the rise in sociological and economic research make hypotheses at the level of collectives common in education.

The habits of the psychologist and biologist do not fit research on classroom instruction. Rats receiving a drug or placebo are properly considered to be independent subjects; what one rat does has no effect on the score of the next (unless the experimenter somehow introduces correlated errors). Students in a class, however, do not provide independent evidence.

Typically, persons within a class are more alike at the outset of instruction than persons randomly sampled from the relevant population. Certain adventitious common experiences during instruction depress or raise the scores of many of them -- a flu epidemic, or perhaps a wave of enthusiasm. What the class experiences goes beyond the treatment specified by the experimenter. There are unintended treatment[1] variations in the experiment on rats also, but the design tries to ensure that no two rats experience the same variation. In the classroom where variation is common to members of the class, the entire class provides a single observation on the effect of the treatment. Some writings on educational statistics (e.g., Peckham, *et al*., 1969) warn against taking the number of students in the experiment as the basis for evaluating degrees of freedom, on the grounds that this gives an unjustifiably small estimate of the sampling error. The warning they do not give is that analysis on individuals often looks at the wrong question.

There is a literature in sociology that warns of the importance of choosing the right unit, and most of that literature too has perceived the question as one of analytic procedure.

[Sociologists (and political scientists, economists, etc.) work with censuses and other public records compiled for some aggregate unit such as a county or an industry. If one wants to know how reliance on public libraries relates to level of education, he may find circulation figures available for each local library system, and educational statistics available for census tracts. Then by combining census tracts that more or less match the service area of the library system, he is enabled to correlate book circulation with education. A famous paper by Robinson (1950) -- echoing E. L. Thorndike (1939) -- warned against interpreting such correlations as if they described relations at the individual level; over the years, a rather sizable literature has reiterated or modified Robinson's warning. I shall later review some of the current ideas, but I do not concern myself with the problem of unavailability of data, which motivated much of the original work.

---

[1] For notes see end of section.

The great majority of sociologists who deal with data at two levels have carried out essentially the same analysis at two levels, or have mingled measures on units and subunits in the same calculation. Thus, a sociologist who had full data might enter into the same regression equation an individual's use of the library, his education, and a measure of the size of the community library, assigning that same value to all residents of the community. The statistical analysis then pools all the individuals, without regard to community boundaries.

I shall propose that in a large class of educational studies, and probably in many other studies of social services, the more reasonable analysis is to relate variables within groups (schools, communities), and then to analyze group-level variables across groups. Whereas most sociologists have related Y to X and to the group mean $\overline{X}$, I propose to relate $\overline{Y}$ to $\overline{X}$, and $Y - \overline{Y}$ to $X - \overline{X}$. This reformulation changes the Gestalt of substantive findings.

Only recently have sociological writers pointedly recommended separate examination of between-group and within-group statistics, though casual references to or illustrations of such analyses appear here and there in the literature. Alwin (1975) has now recommended this decomposition as a superior way to examine composition or context effects, and Firebaugh's (1975) theoretical paper on aggregation bias appears to be in close harmony with this report, insofar as it overlaps. As far back as 1969, Slatin analyzed relations of delinquency to other variables within areal units of a community, and between areal units. In one of the most recent studies of "school effects", Hauser, Sewell, and Alwin (1974) make use of an overall regression analysis, a within-schools analysis, and a between-schools analysis; but they use the analyses to examine somewhat different aspects of the data. Minkowich, Davies, & Bashi (1976) have analyzed the "little Coleman report" on Israeli schools by means of a systematic separation of between-school and within-school relationships.

In writings on "the aggregation problem" or "the units-of-analysis problem", the investigator is presumed to be interested in how one variable depends on another (which may or may not have been manipulated). Writers prior to Firebaugh have discussed whether analyzing means of classes or other groups of subjects is an acceptable substitute for analyzing scores individually, and vice versa. In these discussions, the variables at the group level are "aggregates" of measures on individuals. The choice of unit of analysis has a considerable effect on correlations, regression slopes, and within-treatment variances. Though it has little or no effect on unadjusted within-treatment means, the unit of analysis can make a difference in the estimate of a covariate-adjusted treatment mean, when persons or classes have not been assigned to treatments at random or when the number of independent assignments to treatment is small.

The Abt Follow Through report. The confused state of the art and the importance of the problem are displayed in the analysis of Follow Through Planned Variations data by Abt Associates (Cline, 1974).

The difficulties recognized there and the difficulties not recognized there foreshadow most of my concerns in this report. In this multimillion-dollar study, over a dozen sponsors set up FT groups, each group using whatever model of compensatory education the sponsor advocated alongside an NFT control. Control schools were only roughly comparable to the experimental schools the sponsor used. Comparability of samples across and within sponsors was so poor that Abt analyzed data of each sponsor as a separate quasi-experiment. (I consider this sound; I have doubts about an overall analysis in an appendix that considers sponsors simultaneously.) The fact that the Abt consultants included methodologists prominent in educational evaluation leads me to think that the analysis does reflect the state of the art.

Three analyses were considered: individual, class, and school (each within sponsor). The individual analysis started, in effect, by punching one card with the data for each child. The sponsor's whole batch of FT and NFT children was then run through an ancova program, to reach a number described as the adjusted treatment effect. The school analysis was the same, except that there was one card per school, with pupil averages on variables replacing individual scores. The class analysis was similar, with one card per class, all classes within a treatment being pooled in the analysis.

Abt used different variable sets for the three analyses. Abt chose 18 covariates, but only 11 of these entered the individual analysis and 12 entered the school analysis. The global variable Southern/Western/Other Region, for example, could have been punched in the cards for pupil and class, if it was worth considering at the school level. The school-level aggregate Percent Minority could have been represented at the individual level by a Minority/Nonminority code. (But very likely Abt was not supplied that datum.) Conversely, the individual variable Preschool Exp./ No Preschool Exp. could (and I think should) have reappeared as an aggregate. (Some part of the differences in findings from the three analyses arose because data were missing. E.g., children counted in the class aggregate on certain variables were omitted from the individual analysis because their scores were incomplete. Loss of data is a complication, but probably not the main source of confusion.)

In my opinion Abt was correct to emphasize the school level in its summaries. Treatments were assigned to schools, and no doubt program delivery varied from school to school. Abt, however, feeling that the rationale for choosing a level was weak, offered the three analyses as a "cross-validation". It is not, of course, anything of the kind; the three analyses are in no way independent, and they ask different questions. Abt did indeed state that the analyses asked different questions, and that an aggregate variable is a different variable from the disaggregated variable that generated it. And yet, said Abt, if the three analyses give similar estimates of the treatment effect, the result can be accepted with "enhanced confidence". To be sure, if a critic is disappointed by the finding that the school-level analysis reports, he may claim that analysis at some other level would give the result he would like; presenting all three analyses disarms such a critic. But it is a mistake to regard the three analyses as equally relevant and equally legitimate.

The results of the three analyses did not agree. Sponsor 3, Arizona (pp. VI-66ff.) provides a striking example. Let us confine attention to effects on the Wide Range Achievement Test (WRAT). The first glaring discrepancy appears in the unadjusted treatment effect. With the posttest mean expressed in raw units, the differences (FT minus NFT) are

Pupil	Mean diff. = +1.37	N = 317 FT, 265 NFT	Pooled s.d = 12.8	t = not given
Class	= 3.07	= 38 FT, 26 NFT	= 7.2	t = not given
School	= -0.15	= 20 FT, 21 NFT	= .8	t = -.6

It is to be expected that s.d.'s will be larger at the individual level. It might be expected that means and mean differences will be the same except for such perturbations as missing cases introduce, but they are not. (Discrepancies seem much larger when the Abt charts display each mean difference divided by the corresponding s.d.!) My only guess as to the reason for the discrepant means is that the one-card-per-class and one-card-per-school techniques of calculation weighted cases differently.

I agree with the decision to calculate *t* at the school level only. I do not agree with the decision to test the unadjusted difference, however.

Adjustment changes the picture. The mean differences become

Pupil Mean diff. = 0.36 Change = -1.01
Class = -1.47 = -4.54
School = 0.24 t = 1.37 = 0.39

The adjustment, then, reduced the effect in two analyses, as it should if the FT sample was superior to the NFT sample at the outset. But it increased the effect at the school level, which could only happen if the NFT sample was better at the outset or the regression slope -- positive at the class level -- changed to negative at the school level! At least one of the adjusted analyses must be seriously wrong. In fact, it can be argued that none of them is of much value. The pupil-level analysis and probably the class-level analysis are theoretically inappropriate; and the number of classes or of schools is too small to determine adequately the regression coefficient on which the adjustment is based.

Abt wisely did not test significance at the two lower levels. Many if not most investigators would have done so, if only because the larger N promises a higher significance level.[2] The most serious question to be raised about Abt's significance test is whether it is meaningful.

In a generalized regression analysis loosely comparable to analysis of covariance, 12 regression coefficients plus two constants were fitted to the 41 school means on the covariates. Multiple-regression coefficients are notoriously unstable in small samples; if the coefficients change, the adjustment is likely to change dramatically when the groups are dissimilar to begin with. It is advisable in general to distrust any one regression coefficient when predictors are correlated, even when samples are large. The treatment effect in this study is literally calculated as a thirteenth regression coefficient. I suspect that in a quasiexperiment like this uncertainty regarding the adjusted treatment effect in the population is much larger than the conventional significance test indicates.

The Abt group went one step beyond ancova. Theirs is one of the rare analyses that takes seriously the many warnings in the statistical and psychological literature about fallible covariates. The fallible covariate most likely underadjusts, hence disattenuation is vital in a nonrandom experiment. Abt does disattenuate the adjusted treatment effect in the pupil-level analysis and so arrives at one final "true score adjusted treatment effect". A value of 0.09 replaces the pupil-level "adjusted effect" of 0.36 for WRAT with Sponsor 3. (In other instances, the change is sometimes an increase and is sometimes a change of sign.)

How Abt disattenuated is a mystery. Abt correctly states that the only sound correction method available in 1974 was limited to the study with a single covariate. Yet the analysis they disattenuated was a multiple regression with several fallible covariates. It seems likely that they used one of the unacceptable techniques in circulation in early 1974. Cronbach, Rogosa, Floden, and Price (1976), building on an unpublished paper of Keesling and Wiley, have recently put forth a correction for the multivariate case. Abt might, of course, have hit upon this method.

In any event, the point to be made here is that aggregate data again spawn confusion. Abt corrected only the individual analysis, arguing that class and school data are much more "stable" and in need of no correction. Later we shall see, however, that group regressions may be just as fallible as individual ones. The standard error of measurement of a group mean (with pupils fixed) is small, but the coefficient of generalizability for the group means (which enters the disattenuation formula) may be lower than that for individual data. Class and school analyses of covariance ought to be disattenuated when assignment is not random.[2a]

The need to disentangle effects. Only chaotic debate can result from program evaluations in education until the present confusion about units of analysis is dispelled. The issue is not really one of inference from sample to population, as the infrequent treatment of the issue in statistics texts suggests. And it is not usually one of "substituting" analyses of aggregates for analyses of individuals. Conflicting if not wholly incorrect descriptive results in the Abt sample are the root source of confusion.

Analyses at the group level and the individual level give conflicting descriptive results because they bear on different substantive questions. The investigator who "wants to know the relation between two variables" is not asking a clear question until he tells whether the group or individual relation is the one of interest. The investigator who proposes to partial out certain influences has to specify which relations he intends to remove -- and he had better know why! Some social scientists have recognized that the problem is less one of choosing the right analysis and more one of asking the right question (Dogan & Rokkan, 1969). Scheuch's (1966) exposition -- of how the choice of unit depends upon the theoretical question in hand, and of how the evolving theory takes shape and power from the choice of units once it is made -- is outstandingly complete and eloquent. But even Scheuch is concerned with the choice of units, instead of with the problem of separating between-groups from within-groups effects. ¶ Duncan, Featherman, and Duncan (1972) do have a clear discussion of what is to be gained from such separation, an argument faintly foreshadowed in the marvelously lucid pioneering work of Duncan, Cuzzort, and Duncan (1961).

[Insofar as relevant experiences are associated with groups there are two matters to consider: between-groups relations and within-group relations. The overall individual analysis combines these, to everyone's confusion.

A distinction between aggregate and global data is sometimes made (but not in a consistent way). I shall define an aggregate datum as a simple composite (count, average) of individual characteristics such as per capita income, sex ratio within a school, mean reading level, or percentage of dropouts. Global characteristics are those associated with the collective that are not operationally divisible over individuals, e.g., the per-pupil school budget, the age of a school principal, the size of the school library, the fraction of meetings of a class that are devoted to discussion. A count of a characteristic on which individuals do not vary within groups (e.g., population in an areal unit; sex in sex-homogeneous intact groups) is classed as a global property. The distinction is unimportant, since the two kinds of variables are to be analyzed in exactly the same way. The only real difference is that aggregate variables confuse interpreters, who are inclined to regard the aggregated and disaggregated data as alternative representations "of the same variable". Except in pretest measures on newly assembled groups, they are not (see below).

The interplay between aggregate and individual phenomena can be illustrated by considering the proportion of college-educated in a community. An industry needing an educated labor force is attracted to the area. Then the probability that a person will work in this industry is not merely a function of his individual level of education; it is a function of the educational level in the area where he resides. Causality is equivocal, since the industry, once established, attracts persons with suitable education into the area.

This example draws attention to a point insufficiently emphasized in the literature on aggregation effects. The aggregate variable often represents a different construct from the individual-level variable. A particular relationship might happen to have the same form and parameters at both levels, but even if both relations were described by (say) $Y = 2X + 3$, the relations are rarely "the same". The aggregate $\overline{X}$ and the individual X are different variables; ditto for Y.[3] That the individual is college educated indicates a good deal about what he would be inclined to purchase or what jobs he would be capable of holding. The aggregate college education in the community not only describes an aggregate market and an aggregate employee pool; it says a good deal about what goods and services probably are well supplied in the community (pediatricians? art movies? books? brokerage offices? etc.), and a good deal about the kinds of jobs offered. The aggregate construct enters into a network of relations describing properties of groups (global as well as aggregate properties). It is true that a college graduate is more likely to live in a community where the proportion of college graduates is high. But inference from his individual education to the probability that a choice of pediatricians is available to him is a weak inference, mediated first of all by the characteristics of the group. His individual probability of knowing of multiple pediatricians -- when they are in the community -- does depend on his own education. Instead of considering group and individual relations as alternative versions of the same information, I propose to regard them as statements about different variables, even when the variables originate in the same operation.[3a]

In educational research, practical considerations sometimes suggest that one level is more relevant than the others. The State of California, for example, conducts a testing program whose main function is to inform local district boards how adequate the achievement of pupils in their school system is. The district mean in reading is presented alongside a regression estimate of the expected reading mean. In 1971-73 (for example), the variables given greatest weight in predicting ~~Grade-6~~ reading in unified school districts were an index of family poverty, per cent college educated, and per cent Spanish surnamed. These variables were all aggregated to the district level, and districts were taken as the unit of analysis. This is logical.

The State also reports scores school by school, and compares the school score to a regression estimate of each expected school mean.

There is no a priori reason for the raw-score regression weights for districts to give just predictions at the school level. The State might form a school-level regression equation, entering all the individual schools into the calculation. But this is less logical than a two-step operation that predicts the district mean and the school's deviation from the district mean. The procedure permits assigning one weight to per-cent-college-educated at the district level, and another weight to the school percentage expressed as a deviation from the district percentage (and perhaps a third to the product of the two).[3b]

"Choose the one unit that fits the decision" is an inadequate rule. In a seminar discussion of this report one person suggested that when policy makers want information at (say) the school level, this immediately settles the question of units of analysis. I do not think so; analysis with "school as unit" is not the same as analysis of districts and schools within district. But, in a hierarchical analysis, the results at two or more levels can be packaged into a statement that addresses the question in the decision-maker's mind.

Another example comes from evaluation research. Suppose that an educational innovation will be installed -- if at all -- on a school-wide basis. To decide for or against it one may need to know how student ability influences outcomes. The question can be posed in terms of individual or school characteristics (e.g., the mean ability score of the student body, the range of ability scores). The administrator's question appears to be, In the presence of what school characteristics does this innovation provide cost-effective results? Only if there is a live possibility of reassigning students among schools, or of assigning the students within the school to different treatments, does the decision about adoption rest on individual differences.[4]

This paper examines how to analyze so as to disentangle effects at two (or more) levels and how to interpret both sets (or all sets) of findings.

Psychological bias and sociological bias

As Matilda Riley (1963, pp. 707ff.) has said, it is natural for psychologists to think in terms of individuals and for sociologists to think in terms of collectives. Not only is the psychologist's theory in that form, but the experimental tradition has always looked on the single animal or the single human subject as a biological organism responding to an objective, manipulatable world. Experimental research, even in social psychology, has consistently formulated propositions about a condition that can be imposed "uniformly" on all subjects, as if they were being run one by one in an experimental cubicle. This language has been carried over into evaluation studies and research on classroom learning.

In psychology, units of analysis have received appreciable attention only in connection with laboratory studies of learning. A number of papers (e.g., Estes, 1956) have discussed the fact that "group learning curves" -- i.e., curves fitted to group averages on successive trials -- have little in common with individual learning curves. In particular, a group curve showing gradual learning may actually be a composite of individual curves, in each of which "sudden" learning occurs. Insofar as this discussion has been influential, it has reinforced the psychologist's wish to avoid aggregation.

Scheuch (1966) discussed similar individualist and collectivist biases as they have appeared in economics (and, incidentally, in political science). The attempt to develop theory by combining individual preference or demand functions appears to be the exact counterpart of the psychologist's attempt to combine individual learning curves, save that combining works out badly for the psychologist and analysis at the individual level works out badly for macroeconomics.

A debate in educational psychology. The conflict between this orientation to the individual and an orientation to the group seems first to have been aired in an educational context in 1967 (Wittrock & Wiley, 1970, pp. 271 ff.). In that symposium on evaluation David Wiley stated that the appropriate unit of study in educational evaluation is the collective -- class or school -- rather than the individual. (Today, he would not emphasize one unit to the exclusion of the other.) Wiley was challenged by Benjamin Bloom, who insisted that it is pupils the school teaches. Pupils react as individuals, and the effects on them should be the focus. The instructor and psychologist, Bloom protested, are too often pressed to investigate the wrong question just because it fits into a rationale the methodologists find comfortable. Wiley properly retorted that he had been speaking as a substantive specialist on education, not as a statistician. Upon saying that, he was attacked by Robert Glaser for "ignoring the existence of a discipline called the experimental psychology of learning".

Glaser judged it inappropriate to seek conclusions about classrooms. Effects in the classroom are an aggregation of effects of environmental arrangements on individuals. With a sufficient understanding of the laws of individual learning as compiled in experimental psychology, one would be ready to design environments. A bit later Glaser said, in echo of Bloom: "It is still true that no one has ever taught a class. You teach an individual in the context of a class, but no one has ever taught a class. It is impossible to teach a class. You teach individuals whose behavior changes.... The class is a convenient artifact so that the teacher can reach one student." Against this we can place one of Wiley's final remarks, pregnant for this report: "When we talk about the effects of a treatment on the classroom, we are talking about something fundamentally different from the effects of the treatment on the individuals in the classroom."

Glaser's position does not appear to be tenable. In principle, an adequate account of the laws of learning at the individual level would indeed predict response to any environment, just as in principle an adequate understanding of physical forces at the molecular level would account for the durability of a bridge. The laws that describe learning, however, have to be interactive laws that take into account both the characteristics of the individual and of the setting (Cronbach, 1975). Many of those interactions (e.g., effects on the student of the abilities of the other group members) can only be studied in the group context. That is to say, parameters describing the group have to be written into the "laws of learning." Such relations can only be detected through research on groups of particular kinds (Putnam, 1973).

Debates within sociology. Just as the psychologist prefers to see individual causation wherever he looks, many a sociologist envisions group-level causal processes wherever he can. Aggregate variables have been of particular interest to those sociologists investigating social-psychological processes. The investigators at the Bureau of Applied Social Research at Columbia, and their disciples, have pursued studies of "context effects" with considerable enthusiasm. The central idea is that one's actions and decisions depend not only on his individual characteristics but also on those in his reference group.

[Among reports of context effects or alleged context effects, the one best known to educators is that of Coleman et al. (1966). It was argued there, on the basis of a regression analysis, that a student's achievement and aspirations increase if he is in a student body that is strongly motivated.

Allan Barton (1968) attacked those sociologists who processed data at the individual level, as a prelude to a description of some of the causal models that could be used at the group level:

> For the last thirty years, empirical social research has been dominated by the sample survey. But as usually practiced, using random sampling of individuals, the survey is a sociological meatgrinder, tearing the individual from his social context and guaranteeing that nobody in the study interacts with anyone else in it. It is a little like a biologist putting his experimental animals through a hamburger machine and looking at every hundredth cell through a microscope; anatomy and physiology get lost, structure and function disappear, and one is left with cell biology.

Barton went on to point out that to reduce sampling error the pollster scatters his interviews widely and thereby loses the opportunity to look at behavior in, for example, neighborhood clusters.

Representative of reports of context effects is a study by Bowers (1968) in 99 colleges. Students were asked, for example, if they disapproved of drunkenness and if they had been drunk. The percentage of drunkenness (or cheating, etc.) was crosstabulated (see Barton, 1970) against individual approval/disapproval, within colleges where (for example) the disapproval rate was high. The persons who as individuals approved were less likely to have gotten drunk if the majority in their college strongly disapproved. Hauser (1970b) pointed out that Bowers was in effect entering the group mean $\overline{X}$ and the individual attitude score X into a multiple-regression equation to predict behavior, and then claiming the positive weight for $\overline{X}$ as evidence for a context effect.

Robert Hauser has spearheaded an opposition group within sociology. His 1971 monograph reviewed the literature to that date and challenged those who had tried to show context effects:

> Contextual analysis is based on a misunderstanding of statistical aggregation and of social process which is

> rooted in the identification of differences among groups with the social, and differences among individuals with the psychological. [p. 13]

Bowers' two-variable analysis is of just that character (Hauser, 1970b). Hauser went on (1971, p. 46) to argue that the usual interpretation given the Coleman report is indefensible. Those who are conservative regarding causal interpretations typically refer to "compositional effects", a term apparently introduced by Davis, Spaeth, and Huson (1961).

In an oft-cited paper (1970a), Hauser challenged his fellow sociologists just as Wiley challenged the psychologists. Hauser contrived a demonstration of a context effect: that educational aspiration of students (within either sex) rises as the proportion of males in the high-school student body rises. For the sake of heightening the drama, Hauser went on to propose social policies that would hold down the proportion of females receiving a high-school education. Then he demolished the claim for a context effect by reinterpreting the global sex-ratio variable as a proxy for such aggregate variables as IQ and social status. The groups with high ratios also were higher in the proportion of high IQs and students of high status.

Hauser's argument is essentially about specification error. If one relates the dependent variable to only a fraction of the initial variables at the individual level that contributed directly to the effect (or that contributed to the allocation of persons into groups), this is equivalent to using an inadequate covariate to adjust scores in a quasiexperiment. Only if an ideal adjustment is made (Cronbach et al., 1976) will one properly evaluate the effect of groups as such.

Barton (1970) challenged Hauser's argument and Hauser (1970b) replied. The debate continued in a paper by Farkas (1974) and a rejoinder by Hauser (1974). The several papers cite earlier arguments for and against contextual interpretations. It is unnecessary to restate the several positions, particularly since I am advocating a kind of analysis not discussed directly by the others. It may be useful to restate the essence of Hauser's position as I understand it. The heart of the matter is a rule of parsimony; if most of the variance can be explained by individual-level relationships, there is no need to invoke a contextual explanation. Thus, where Bowers gave X and $\overline{X}$ equal status in his regression, Hauser considers it appropriate to calculate regression weights for X and $\overline{X} \cdot X$. Since X and $\overline{X}$ are correlated, this procedure allocates most of the predictable variance to the first predictor. (My proposed scheme is similar to Hauser's save that it fits weights to $\overline{X}$ and $X \cdot \overline{X}$ -- which equals $X - \overline{X}$.) Hauser does not deny the possibility of causal effects at the group level, but he places on them the burden of proof. Moreover, and his point is one that no writer of the 1970's would deny, any serious claim to a group-level causal effect ought to be supported by tracing it to observable intermediate processes. Simple pre-post correlations or regressions do not carry much weight in a discussion of causes today.

The terminology of the sociological debate has been an unnecessary source of confusion. I suggest that three kinds of relations are worth distinguishing:

1. Demographic effects. The groups examined have, as groups, no causal influence. But the groups differ on certain precursors of the outcome variable of interest. Processes at the individual level would generate outcome differences between the groups. This is Hauser's preferred explanation for observed effects at the group level. One might speak of "composition" effects, but there are ambiguities in the term. If desegregated schools create outcomes

unlike those the same students would have had in segregated schools, this is a consequence of student body "composition". While "demographic" is open to the same construction, I think it can serve as I have defined it.

2. Group-caused effects. Outcomes for a given individual depend on the group he associates with or the setting in which his group works. This includes "context" effects that arise from peer influence, and also "school" effects that arise from particular curricular offerings or other nonpsychological causes. Insofar as the events in the desegregated school modify outcomes, the effect is "group-caused". To be sure, a new curriculum is not caused by the group, but it is a cause that affected the person because he is a member of the particular group.

3. Arbitrary aggregation effects. The relations listed above apply to the study where groups are observed over a period of time and changes are to be explained. Grouping is sometimes imposed on a body of data after the effects have been produced. This happens when survey data on, for example, race and unemployment are aggregated to the level of, say, the county. Insofar as the basis for aggregation correlates with either or both the variables of interest, statistics at the aggregate level may differ from the corresponding disaggregated statistics.

As we proceed it will become increasingly evident that, from data on X and Y alone, it is impossible to establish which of these classifications a phenomenon falls into.

Units of analysis? of treatment? of theory?

Abt, handed data to process, saw the question as one of units "of analysis". At an earlier stage in the Follow-Through evaluation, however, the question had been faced as a choice of units of design, i.e., of sampling and of treatment. The sponsor was instructed to identify schools in which he would install his FT treatment and similar schools to be NFT controls. This only crudely approximated a process of formal sampling and random assignment, but it did identify the school as the unit to which the treatment would be administered. (The plan actually called for treating just a few classes per experimental school, ignoring the others.)

A structurally different decision was made in designing the Performance Contracting experiment (Ray, 1972). Districts were chosen as before, somewhat arbitrarily, and two schools with disadvantaged pupils were selected within the district. One of these went into each treatment. The district was a sampling unit -- given an intent to generalize the results into national policy. The school, however, was the unit of assignment, hence of treatment.

Someone might challenge this terminology by describing a study with the same design where the treatment was a vaccine administered to each experimental student individually, with a placebo administered in the control school. Individual injections or no, I still see the treatment unit as the school. The design equalized district factors over treatments, but it confounded school factors with the treatment. If this design was consciously preferred to a split-school design, the justification must have been interest in some social effect (e.g., spread of the disease in an inoculated *community*).

It is possible for the unit of sampling and the unit of treatment to differ in other ways. One might sample individuals and assign them to classes individually, and then assign classes to treatment. Then the unit of treatment is clearly the class. Conversely, one might sample classes and then assign individuals from the classes to one or another independent treatment.

This brings us to the unit "of theory". The choice of design is often constrained by practical matters, but the rationale for the design ought to come from theory. Theory need not be grand and abstract, but it does state a question in general terms. The wrong design may examine too broad or too narrow a phenomenon. Federal support for Performance Contracts was entertained as a national policy, but it was surely anticipated that each district would decide whether to enter such contracts. Hence a contrast of experimental and control districts would have been sensible.[5] If the thought was that PC, once adopted, would become mandatory for all districts in the nation, the logical experiment would, on its face, be a period of nationwide trial. The contrast group could be another nation or the same nation in the pre-experiment period. The notion of taking the nation as the unit of treatment may be dismissed if theory says that every effect is mediated locally. In some contexts no such claim would be made. America's "noble experiment", the Eighteenth (Prohibition) Amendment, could not have been evaluated by studies of prohibition at local option.

To define a unit of theory is to argue that there are boundaries in the social space which mark off entities that have properties of their own. Just how to identify "entities" or "systems" for scientific study, where object boundaries are not apparent to the eye, is a question of long standing in many fields including sociology (D.T. Campbell, 1958). Some social entities appear to be good subjects around which to build theory; they cohere, and their members undergo common experiences. Other groupings (e.g., by first letter of one's name) have no more than momentary power to produce a common effect on the group members. Groups that are real for some purposes (e.g., college majors) are unlikely to be the groups around which some other aspect of behavior (e.g., social life) is organized. Groups that are interconnected in some respects, hence part of a larger system, may function as independent systems with respect to certain phenomena.

Analysis at the level of the collective is likely to have no justification in science or in policy studies unless the collective is in some real sense a carrier of an effect. Shively (1969, p. 1184; his italics) warned against calculating ecological correlations, and presumably would warn against regressions also, "unless the theory with which we are working conceives of the aggregations we are using as real entities, *for which no other type of aggregation can readily be substituted*." In educational research it does seem reasonable to think of classrooms and schools and districts as having real enough effects. To analyze at the group level seems to invite no greater penalty than the disappointment of looking for a group-level effect and finding it absent. In other kinds of research, the social fabric may be so seamless that no unit of theory can be readily defended. Then some model other than that of members-nested-in-units may be required.

Hannan (1971) considers that the so-called aggregation problem in sociology (and economics[5a]) arises as much from the units of theory as from the units of aggregation for analysis. (Where, as is usual, sociological and economic data are collected naturalistically, no question of unit of assignment arises, and Hannan does not concern himself directly with sampling units.)

Macrosociology and macroeconomics seek generalizations applicable to large collectives, whereas microtheory seeks to generalize about processes occurring among small units (e.g., social participation or purchasing behavior of the single family). Propositions at the two levels may be cast in terms of the same construct (e.g., per capita income). Some sociologists -- Hannan points to Parsons as arch-example — expect "homology", with the same relations emerging at all levels once the right set of variables is identified. Others, including Hannan, expect the micro path coefficients linking homologous variables to differ from the coefficients generated by macrodata on the same sample. He sees the ultimate problem not as picking a unit of theory but as of developing a "between levels" theory of aggregation processes, to permit reductive interpretation of macro data and aggregative interpretation of micro data. Micro, macro, and aggregation relations together constitute an ideal theory for Hannan.

What social scientists have generally seen as a problem of data analysis has a striking correspondence to a major issue in the philosophy of natural science, reductionism. Daniel Bell (1975) discusses the attitudes that physicists, in particular, have taken to the proposition that relations need to be developed in an integrated manner so that one can read upward from subnuclear processes and downward from phenomena on the human scale and larger. How close Bell is to our concerns is indicated by the fact that at the outset he quotes J. S. Mill to illustrate "the 'naive' formulation of the issue":

> Human beings in society have no properties other than those which are derived from and may be resolved into the laws of the nature of individual man.

The need to move back and forth between levels of aggregation is minimal for the physicist, Bell suggests,because the energy that goes into processes within atoms -- for example -- is some orders of magnitude below the energies that go into the kinetic motion of gas molecules at normal temperatures. The proposal to look at each level in turn in social research cannot use that rationale; the energy in individual transactions must be of the same order of magnitude as most contextual effects arising from the class. Apart from that, my proposal to examine class-level relations and then to examine individuals-within-class is in striking parallel to what the physicist does. Having studied molar gas laws to his heart's content, he turns to the study of forces binding atoms within the molecule. But he seeks conclusions about atoms within a pure gas, not a conclusion about atoms without regard to molecular context.

I can see the benefit to be gained -- in principle -- from Hannan's integrated theory. Suppose our question is, What will it do to students' life chances if we require passage of an achievement test before allowing them a high-school diploma? A local ruling, a state ruling, or a national ruling would have different effects, and a theory of the type Hannan has in mind could hypothetically forecast them, without experimenting at all three levels in turn. I have no faith that social scientists can attain such powerful theory (Cronbach, 1975). If I am right, it is necessary direct one's inquiry to whatever level is most pertinent to the question of theory or practice of most immediate concern.

Units within hierarchies. Almost all previous writers have spoken of the contrast between analysis of elements and analyses of collectives, Abt and Hannan being recent examples. My plan of attack is instead to pick one level of collective and examine (a) relations between collectives at that level and (b) relations of elements *within* collectives (rather than relations of elements without regard to the boundaries of their collectives).[6]

Almost all my argument will be confined to two stages -- e.g., pupils within classes. I shall consistently treat a measure on the smaller unit as a composite of a mean for the larger unit and a deviation from that mean. (E.g., mean age of class, and pupil age minus class mean.)[7] There is no logical difficulty in extending such a series of components over pupils-within-classes-within-schools-within-districts. With a dependent variable at one level, all components of independent variables associated with that level and higher levels may enter the analysis. (Also, a statistical index derived from a component at a lower level may become an independent variable. E.g., the s.d. on a predictor of pupils within a class may be used to account for class mean differences on the dependent variable.)

The same principle could be applied in the reverse manner, the class mean being a composite of pupil score and class-mean-minus-pupil-score. Then the dependent variable at one level is explained by components at that level and lower levels. For some studies this may be more appropriate than downward decomposition. The chief difficulty is that the deviation score is correlated with the *lower*-level score, which complicates analysis and interpretation. (The score $\bar{X} \cdot X (= \bar{X} - \eta^2 X)$ has a zero correlation with X.)

This upward decomposition was central to Blau's (1957) definition of structural (context) effects. When Hauser restated Bowers' question in terms of regression weights for X and $\bar{X}\cdot X$, he was proposing a similar upward decomposition. This formulation seems to be the one that springs to the mind of those sociologists who choose not to express relations directly in terms of X and $\bar{X}$. Perhaps this follows from the obvious causal principle that $\bar{X}$ arises from X and from the sense that a context effect is something added. A reference-group perception, however, might easily operate causally in terms of $X - \bar{X}$, as is seen in Meyer's hypothesis (1970, p. 63) that a student's judgment of his own ability -- which affects his aspirations -- arises from his standing relative to his group. I do not argue that $\bar{X}$ is prior to X; rather (p. 3.3), I make X prior to $\bar{X}$ and $X - \bar{X}$. I also partition Y, which has not often been done in the sociological literature. Insofar as I have a causal preconception, it is that $\bar{X}$ often determines what educational activities are offered to a class or student body. But no one causal position fits all studies.

Units in areal analysis. The procedure probably does not apply well to all the kinds of nesting considered in writings on aggregation. Where exposure to treatments takes place in South Sea islands, and the islands are assembled into collectives each of which unites the islands under its own policy, hierarchical analysis applies. This is equally the case with classrooms nested within schools. In the old problem of slicing up a time series, however, months nested within years or biennia are not islands. A price movement is not contained within a month or a year. Areal units similarly flow into one another and, at least in agricultural research, the reporting area corresponds badly to the causal variables such as weather and marketing facilities.

A model of units nested within larger units may be unrealistically simple even in schools. In simpler days, pupils were nested within classes, firms within industries, families within communities. Today, even the 9-year-old may work in a dozen groups and individual settings with several teachers and aides, all in the course of a school day. Similarly, the firm is often a conglomerate, and family members commute and so come under the influence of several communities.

Streuning (in Streuning & Guttentag, 1975) points to the importance, for evaluation of health services, of an ecological analysis. He proposes to divide a catchment area with a population of perhaps 400,000 into 100 units, and to correlate characteristics of the unit with indicators of use of services. His plan calls for reducing a large set of predictor variables and a large set of dependent variables by cluster or factor analyses, followed by calculation of multiple-regression equations relating dimensions from the two sets. The coefficients in these equations would generate hypotheses about ways to increase use of services. Streuning's plan probably represents, at its best, the state of the art in using existing records to understand and improve a social program.

The question to consider here is whether the choice of unit matters. Streuning chose not to use a unit smaller than the census tract, presumably because many data available at that level are unavailable at lower levels. He chose not to use a larger unit because a sample size of 100 or more is recommended for inference from sample correlations. Streuning does not argue that the tract boundaries relate to any gradient of action. Some actions -- say, reducing the income level at which a certain service is given without charge -- are conditional on individual characteristics. Some -- establishing a hot-line phone for pregnancy counseling -- are citywide. Streuning could well have thought more about the choice of unit. The N of 100 units should not be a ruling consideration for Streuning. Insofar as he is seeking policies for this single catchment area he is dealing with a population, not a sample. Insofar as he is seeking theoretical insight he is dealing with a sample of size one (of many catchment areas in the nation). Although it may be distressing not to have some data for units smaller than the census tract, this is not an insuperable barrier to using smaller units; one can assign the value for the tract, or a prorated value, to each of its subunits (say, an apartment building).

Geographical areas can be divided coarsely or finely. The Yule-Kendall computations on wheat and potatoes (p. 2.12) show that correlations change with the unit of analysis. Some correlations will change more than others. If so, both at the factor-analytic stage where he reduces the predictor set and at the multiple-regression stage, Streuning could expect to get different results as he alters the unit of analysis.

As no areal unit can be seen as the unit of theory in Streuning's case, it is uncertain what procedure to recommend. A next step appears to be to collect data that are disaggregated to the greatest degree possible, and to apply the proposed methods of analysis across and within various alternative levels of aggregation. A serious problem for studies of human ecologies -- once we leave the neat hierarchical partitioning of schools -- is how to bound "an ecology". Arbitrary slicing of areas along the lines of large aggregate reporting units defined without reference to the problem in hand seems certain to misdirect thinking.

Notes for Section 1

Page

1.2 [1] In this report, treatment is a general term. It includes controlled and administered instructional or therapeutic interventions, but it also includes variations in services that sprang up without control (e.g., talkative teachers vs. listening teachers). Any service or activity or policy that could in principle be installed deliberately is a treatment. Although many examples in the first parts of this paper refer to treatment contrasts, the theory to be developed considers relations within a single treatment. It therefore applies not only to experiments but to naturalistic studies (e.g., of utilization of educational TV).

Much of the discussion of units in sociology has been concerned with correlations between variables that are present simultaneously (e.g., ethnic and religious identifications). Some of my thoughts about asymmetric relations of treatment to outcome may not fit these studies where there is no manipulable variable.

1.6 [2] In an appendix, Abt did report child-level significance tests, claiming as many as 3800 d.f.

1.9 [2a] Lumsden (1976) takes vehement exception to my advice regarding disattenuation. I should apologize for recommending disattenuation and yet reporting attenuated results throughout this work. The examples are all secondary analyses, and at best I could show the effect of a guessed reliability coefficient on any results.

Page

1.11 [3] This is not true, of course, when groups are formed at random and treated individually. Then any relation of $\bar{X}$ to any variable is nothing more than a "composition" of the relations of X. When aggregation is after the fact, and individuals within an aggregate have been treated independently, the statement may or may not hold. Consider race as the basis for grouping. Income within the race group may be an indicator of successful performance; income as a between-group variable is heavily colored with market inequity. The variable in the total pool is a mixture of the two constructs.

1.12 [3a] Blalock (1964, p. 98) says that when relations of Y to X and $\bar{X}$ differ the relation must have been altered by the entry of certain causal variables at one level and not the other. I prefer to say that $\bar{X}$ and X are distinct variables. The properties of what the physicists call a critical mass arise from the aggregate itself, not some "additional variable". The whole in this case is more than the sum of the parts.

1.13 [3b] In the one trial of such a scheme that we have made, the three analyses generated much the same standard deviation of residuals in a cross-validation, though the regression equations did not weight the variables in the same way.

Page

1.13 [4] Even then, research conducted in schools as now constituted is a poor basis for forecasting what will happen when new assignment rules are adopted.

1.19 [5] From schools contrasted within districts one can generate a difference score between treatments for each district. What appears to be a school-level design is thus capable of being given a district-level analysis. A policy decision that PC should or should not be adopted district-wide hereafter, on the basis of a difference in this study, does require the assumption that the effect in a PC school is the same whether or not comparable schools in the district have PC. The choice of unit for assignment of treatment thus rests on a theoretical proposition. Even the district may not be a large enough unit for adequate evaluation of a policy. The Federal government entertained the idea of encouraging district use of such contracts; but if the designer of the evaluation could offer grounds for believing that payoff would be greater when all the districts in a county went on the contracting basis, then sampling scattered districts would not disclose important data on the working of the policy.

1.19a [5a] There is a substantial literature in economics and econometrics that I have made no attempt to review.

1.22 [6] Analysis of elements within collectives is of course the basic method in comparative studies that take one nation or one school at a time, and then repeat the study in another collective. I know of no instance in which such a comparative study has formally analyzed across collectives as well as within collectives.

1.22 [7] The dichotomous variable such as Black/white becomes π = per cent black for the class, and $1 - \pi$ or $-\pi$ for black or white individual, respectively. Although the two components are linearly independent, the variance of the deviation component is nonlinearly related to π.

2. Units in various research contexts

The problem as seen in research on Aptitude × Treatment interaction

I was brought to face the aggregation problem while R. E. Snow and I were completing a review of the numerous studies on Aptitude × Treatment interaction (ATI; Cronbach and Snow, 1976). The issue in such research is whether outcome-on-aptitude regressions (hereafter, I refer to Y-on-X regressions) have the same slopes within the treatments -- say, within competing teaching methods. Findings on interaction might give the school a basis for assigning a particular student to whatever mode or style of instruction is likely to produce best results for him. The investigator naturally approaches the problem with the psychologist's bias, asking how individual characteristics relate to individual outcomes and hence to choice of treatment for the individual. Conventional thinking about aptitude effects on individuals is not fully applicable, Snow and I now realize. A practically relevant conclusion ought to describe the result to be expected under the usual school conditions. The school teaches students in groups -- even in much "individualized instruction".

Sample size for regression analysis. Investigators conducting ATI studies in classrooms have, with rare exceptions, pooled the data for subjects within a treatment before analysis, ignoring the class grouping. They have taken the individual student as the unit of analysis.

Even in a simple t-test on the outcome in a true experiment, a calculation at the group level is less likely to reach significance than a calculation at the individual level. Assuming uniform group size n, the two t values have approximately the ratio $n\eta_Y^2$. Thus if η_Y^2 has a reasonable value of 0.30 and $n = 10$, the individual t is 3 times the group t. In regression analysis the lack of power is even more serious. (See Section 7.)

Something like 100 degrees of freedom is required to reject the hypothesis that a regression slope is zero when the actual slope is large enough to be of considerable importance -- say, a standardized slope of 0.4. (See Cronbach & Snow, Chap. 3,4.) Consequently, 100 classes (!) must be observed to get a good fix on a between-classes regression. Pooling classes and analyzing individual scores, the investigator claims a large number of degrees of freedom; he is then more likely to be able to report a significant difference between regression slopes. Unfortunately, his significance levels are spurious unless he makes strong assumptions. If classes are the unit of sampling, the number of classes is the natural basis for statistical inference.

The strategy of ATI research (and of much other social and educational research) will have to be modified, once it is recognized that the costs of the usual strategy are nearly prohibitive. Experiments examining the difference between group-level regressions will be uninformative, in the sense that the prior and posterior probabilities of accepting the null hypothesis are nearly equal in a study of reasonable size. Only with the sample sizes attainable in survey research will one find it profitable to assess the "significance" of group-level regressions.

The Maier-Jacobs study. One team investigating ATI did take classes as the units of analysis. Maier and Jacobs (1966) carried out a year-long experiment in many classrooms. Spanish was taught by programmed instruction in 39 elementary-grade classes; 17 by an "orderly" and 22 by a "scrambled" program. Maier and Jacobs analyzed the classroom means on various pretests and outcome measures and reported, among other conclusions, that the outcome means were similar in the two treatments. The between-classes regression slope of attitude toward programmed instruction (posttest) onto IQ was positive when the orderly program was used. (The slope was small because the s.d. of attitudes was very small in the metric used, but the correlation was 0.75.) The implication is that duller classes liked the orderly program less than abler classes. In the scrambled treatment there was an effect in the opposite direction; programmed instruction received higher ratings in duller classes.

Another set of between-classes regressions used IQ as the predictor and an achievement posttest as the outcome. Maier and Jacobs provided Snow and me with statistics on those variables for all cases pooled, from which we could calculate three sets of achievement-on-IQ slopes:

	Orderly	Scrambled
Overall (individuals pooled)	0.53	0.62
Between classes	0.50	0.77
Within classes (pooled)	0.55	0.52

The differences are not enormous, and no sensible comment about significance can be made with a limited number of classes per treatment. Evidently, abler classes pulled ahead of duller classes, most strongly in the scrambled treatment. Second, IQ differences within the class related to outcome similarly in each treatment. This (like many other studies) denied the working assumption of the programmed-instruction movement of the 1960's, that orderly step-by-step progression of instruction would largely erase the effect of IQ on learning.

Harvard Project Physics. The evaluators of Harvard Project Physics, an innovative high-school curriculum, likewise collected data in classrooms scattered over the nation. In addition to individual scores on beginning-of-year and end-of-year tests, the investigators had information on the climate of each classroom, obtained by aggregating questionnaire responses of students. The papers of this project sometimes reported analyses at the individual level (cases from all classroom being pooled) and sometimes reported analyses of class means. The chief report published to date (Welch & Walberg, 1972) analyzed at the class level. The several analyses in this and earlier publications cannot be directly compared because they used somewhat different variables and statistical techniques. Interactive effects were reported. The studies suffered from a number of faults common in research on interactions at that time. Snow and I, after a critical look at the methodology of the studies (Cronbach & Snow, 1976, Chap. 10), were uncertain as to the dependability of the interactions reported. The comparison of main effects, using classes as the unit of analysis, was open to much less question.

Walberg (paper in preparation) has recently reflected on the HPP experience. He comments that the research group received conflicting advice from methodological experts as to the best way to handle the mixture of individual and class data they had amassed. His paper (in its current draft) goes on to list something like a dozen competing modes of analysis, several of which were tried by himself and his colleagues on one or another set of variables. His final paper will be of obvious use to persons interested in this report; but it would be inappropriate to discuss the draft here.

Only late in our work did Snow and I become aware that the interaction phenomenon has to be defined substantively as a between-groups or within-groups effect.
We came at last to see the importance of Wiley's view that response to treatment is not simply an individual-level process.
Group characteristics -- aggregate or global -- may interact with treatments, and they may interact in a different manner than individual characteristics.
Some pages on this theme were added to Chapter 4 of the Cronbach-Snow book in the final stages of writing.

Three kinds of process. To recapture the argument of those pages, it will suffice to consider the regressions of outcome Y onto aptitude X_p (the score for the individual p) and onto the aggregate X_c, the mean of this same aptitude over individuals in class c. At least three kinds of causal phenomena may enter into an observed interaction or an observed regression slope calculated from individual-level data.

There is a sample of classes. These need not have been assembled at random, but the classes are divided at random between two treatments. A common outcome measure is obtained on all persons, with the following hypothetical results:

a. The overall mean is the same in treatments A and B.

b. The regression of outcome on a measure of initial ability or achievement is nearly flat -- say, a slope of 0.2 -- when calculated on all the persons in the A group.

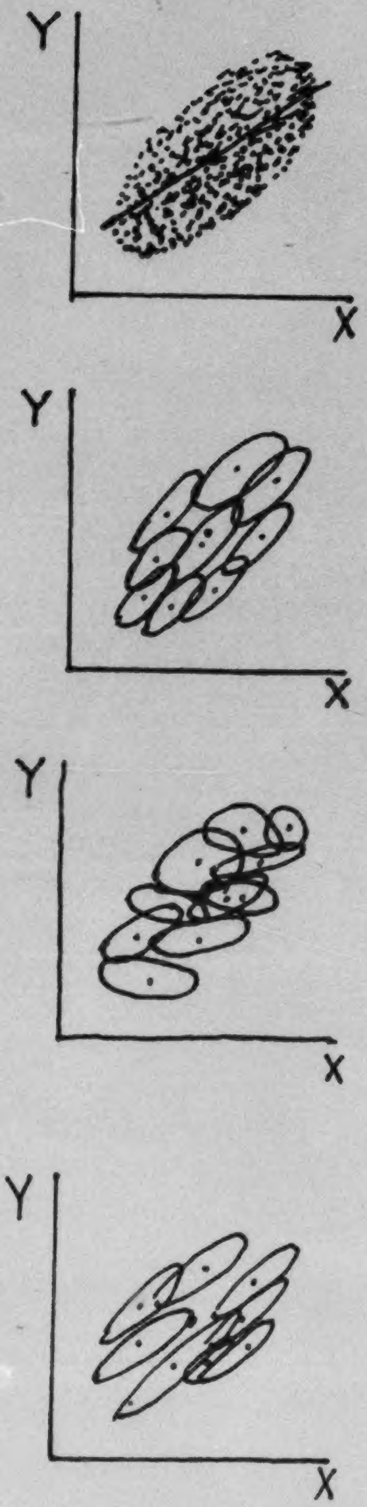
c. In the same metric, the individual-level regression slope is 0.6 in the B group (Figure 2.1). I.e., students with superior aptitude do considerably better in B than students of low aptitude, and considerably better than their high-aptitude counterparts in Treatment A.

Three alternative explanations may be offered:

1. Interaction at the individual level.
2. Interaction at the class level.
3. Interactive effects within the class.

A concrete example will help in what follows. In Treatment A (didactic), history students study immigration problems of the U.S. through textbooks and teacher exposition. Treatment B is inductive; the students examine original documents, newspapers, etc. and work out conclusions through discussion.

at 2.7



(i) Result observed in pooled data

(ii) Regression generated by individual effects, with between-class and pooled-within-class slopes equal

(iii) Regression generated wholly by class-level effects; between-class slope greater than pooled-within-classes slope

(iv) Regression generated wholly by within-class effcts; between-class slope is zero.

Figure 2.1. Alternative ways of generating an overall regression slope of 0.6

The three kinds of effects can be examined without reference to differences in slope (interactions). The prior problem is to explain the within-treatment regressions. How might the steep regression (all cases pooled) in the inductive treatment have arisen? Let us assume that the groups differ at the outset of the study only with respect to X and irrelevant variables; i.e., there are no specification errors. Also, to hold questions of attenuation effects in abeyance, let us assume that X is perfectly measured.

(1) Individual level. Psychologists have regarded regressions of outcome on aptitude as manifestations of individual aptitude, working on lessons delivered to the individual. If a Y-on-X regression is steep, the interpretation has been that the person with a high score on the aptitude test somehow processes material more efficiently or diligently than the low-aptitude student. E.g., the fast reader has a considerable advantage when numerous documents must be scanned and evaluated for relevance. Interpreting the observed interaction as of type (1) implies that the result would be found if students were exposed to the teaching method individually. Panel (ii) is consistent with this interpretation. The within-class regressions depart only by chance from the pooled regression. Any particular configuration of regressions could arise from combinations of two or three kinds of effects, however.

The modern interest in ATI stemmed from a concern with individual assignment in education. Cronbach and Gleser (1957) established a rationale for validating such assignment rules. To justify assigning students to alternative treatments (e.g., to regular and slow sections), it was logically necessary to show that a steeper regression existed in one treatment than in the other. Although the matter was never discussed fully, Cronbach and Gleser assumed that data would be collected by assigning students from the whole aptitude range to each treatment, just as a selection test is validated on a sample from the whole range of applicants. In suggesting that regressions observed in wide-range groups would guide the formation of groups more homogeneous in aptitude, Cronbach and Gleser implicitly assumed that the regression slope reflects the response of the individual to the treatment. His expected outcome in a given treatment was taken to be the same regardless of the choice of persons to be treated alongside him.

(2) Group level. An alternative causal hypothesis is that the level of aptitude in the class as a whole determines the effect of a treatment. Would not a steep slope be found in Treatment B if the source material selected for interpretation in abler classes (as identified by mean aptitude) were much superior to that selected for use by the dull groups? Such a mechanism, triggered by the class average, perhaps serves both the abler and less able members of the able class. Under this hypothesis the richness of the experience depends on the environment, not on the abilities of the students working singly. (See Panel iii.)

A similar group effect might be found if the teacher regulates the pace, forcing the discussion to a penetrating level in the able class and leaving it superficial in a dull class.

(3) Comparative effects within the class. The third possibility

is that the regression slope is determined by effects within the class. Suppose that, in classes using the inductive method B, the ablest students within the class steal the show. They dominate the discussion; they are rewarded for locating materials more rapidly than others, and so are encouraged to redouble their efforts. The duller members of the class, systematically outshone, come to rely on their abler classmates to keep things going (Panel iv). Another possibility is that the typical teacher will habitually interact more with the superior members of the inductive class.

Effect (1) presumes that the outcome is a function of the student and the choice of treatment, not depending in any systematic way on the makeup of the class. Effects (2) and (3) presume that class makeup matters, that two students whose aptitude is at the population mean will achieve differently when one is superior to the classmates he draws and the other is assigned to a group abler than he is. In Panel (iii) a student gains by entering a class where he is below average; in Panel (iv) it is the other way around.

The shallow slope in the didactic treatment A might be explained by the near-absence of all the three types of regression effect. But effects can balance each other. A shallow pooled slope in the didactic treatment may result if one effect is positive whereas the other two are close to zero or one is negative. It is possible for a slope to be negative (e.g., when a teacher concentrates effort on the least able members of the class).

The difference of 0.4 in slopes in the pooled analysis (= 0.6 - 0.2) can arise in principle from an interaction effect of 0.8 at the individual level, an effect of -0.4 (say, 0.2 - 0.6) at the group level, and an effect of 0.0 at the within-group level. Figure 2.2

sketches two out of many possible configurations that yield pooled slopes of 0.2 in A and 0.6 in B. An interaction observed in pooled data from many classes obviously cannot be directly interpreted.

The problem I began with, then, was this: What analysis comes closest to describing separately the interaction effects of the three kinds? And what problems of interpreting the findings arise?

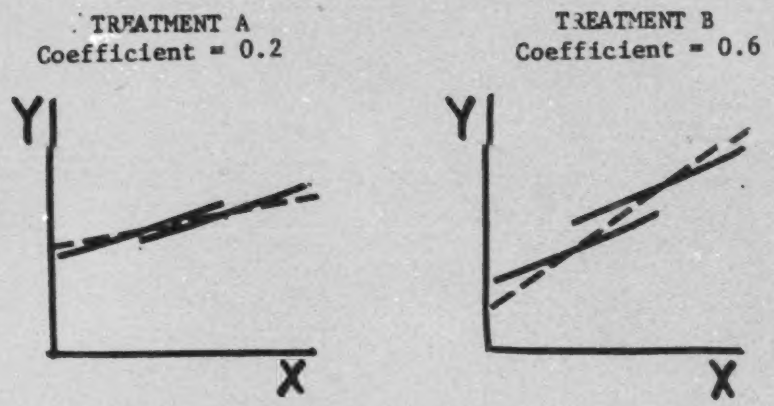
It has perhaps already become apparent to the reader that the problem is not adequately formulated in the paragraphs above. Once a distinction between effects at the individual and class levels is made, it is natural to separate within-classes and between-classes regression analyses. At best, this resolves into two components an effect that has three possible sources. I see no way to disentangle the effects in analyzing data from the usual designs.

Aggregation effects

The sociological and econometric literature contain many papers on what is usually referred to as aggregation bias.

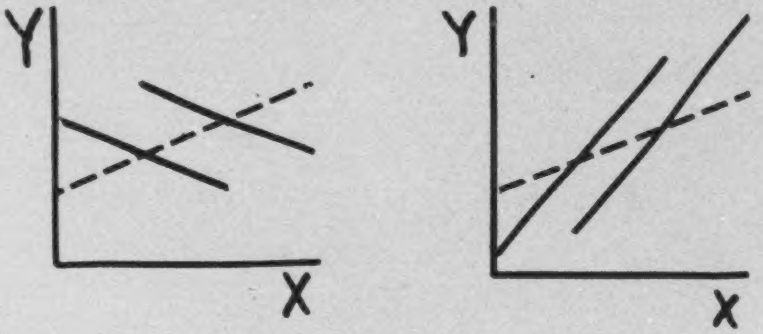
Robinson (1950) set in motion the discussion of aggregation bias in sociology by demonstrating the disparity. Later papers have judged that Robinson was shortsighted to emphasize correlations rather than regression coefficients, but the same issues arise when regression coefficients are compared. Although the literature triggered by Robinson's work is voluminous (see citations in Dogan & Rokkan, 1969, and Hannan, 1971), thinking has not moved steadily forward. Arguments have become more complex, but consensus is lacking. As late as 1971 Hauser could say that "sociologists have not yet fully exploited the insight it [Robinson's article]

at 2.11


TREATMENT A
Coefficient = 0.2

TREATMENT B
Coefficient = 0.6

(i) Difference in slopes at the between-group level



(ii) Difference in slopes at the within-group level

Figure 2.2. Distinctive configurations generating the same overall ATI (The intraclass correlation is assumed to be 0.75.)

provides into the interpretation of relationships at different levels of aggregation." (His p. 11-12.) He went on, in echo of Riley (see p. 1.14) to speak of a "misunderstanding" arising from the view that effects at the group level are sociological in nature, and individual effects psychological ("internal to the individual"). Firebaugh (1975) rejects even today the complacent view of Scheuch (1966) that most of the problems are "intellectually settled".

The original concern for aggregation bias had to do with the effect of arbitrary contiguous grouping. Yule and Kendall (1950, p. 310-11) chose the example of a correlation between potato and wheat yields per acre. It is necessary to use data at some aggregation larger than the potato patch; in the records, counties were the smallest available unit. It seems reasonable *a priori* to group neighboring counties. In a series of correlations for 48 counties, 24 pairs, then 12 sets, and finally 6 regions, the correlation moved up from 0.22 to 0.76. Yule and Kendall properly concluded that a correlation is specific to the unit chosen, but failed to consider deeper interpretations. Some of their *within*-region covariance was part of the overall *between*-county covariance. Aggregation redefines the question investigated by moving certain information out of one variance (covariance) and into another.

Härnqvist (1975, pp. 101-102) offers some educational examples with non-arbitrary grouping. He correlated measures collected in the International Study of Educational Achievement at the individual, school, and country levels. The respective correlations of reading comprehension with a measure of school satisfaction among 10-year olds were 0.19, 0.18, and -0.77 -- a change of direction as well as size. He also shows that it is possible for aggregate correlations to be smaller in absolute value than lower-level correlations. The successive correlations for reading with science knowledge at age 14 were 0.60, 0.76, and 0.54.

Discussion in recent years has centered increasingly on causal interpretation. A number of writers rejected Robinson's emphasis on one correlation as an estimator of another, making the valid point that the two correlations reflect different phenomena or processes. It is the failure to arrive at a clear logic for interpreting the two that continues to plague the field. Thus Patricia Kendall and Lazarsfeld (1955, p. 295) discussed data from The American Soldier where a within-groups regression coefficient was positive and a between-groups coefficient negative. Soldiers who had been promoted gave more positive answers on a question about promotion chances in the Army than men who had not been promoted. Ratings related positively to individuals' actual promotion. This was true for military police and also true within the Air Corps. But promotion rates were higher in the Air Corps whereas the rating on promotion policy was higher in the MP's. That is, the between-group regression of rating on mean actual promotions was negative. Kendall and Lazarsfeld concluded that the group phenomenon reflects shared experience and perceptions and is not just an aggregate of individual data.

Oddly, they abandoned this caution in another instance. Soldiers who chose their own assignments liked their jobs better than others did. Units where choice was commonly allowed were most likely to be rated by their members as good units. So the between-groups and within-groups slopes are similar. Kendall and Lazarsfeld said that the individual relationship "corroborates the result" from group data. This appears suspiciously circular. If the results had disagreed (as they could have), would the writers not then have insisted that the group and within-group data bore on different phenomena?

For the sociologists in the Columbia group, most processes (e.g., generating a certain income level) occur to the person in a group context, and no process truly operates on the individual in isolation. Coleman (1954) urges contextual interpretations because otherwise sociology becomes no more than an "aggregate psychology". It appears that one can interpret an aggregate regression coefficient as causally similar to an individual coefficient only by assuming that group members developed their scores on the variables independently, two members of a group sharing an experience no more often than members of different groups. Moreover, even if Y scores of individuals were generated independently, if groups were formed on the basis of any variable that correlates with Y·X this will produce a difference between the within- and between-groups slopes. (See Section 3.) There can be no general warrant for substituting group data when individual relations are of interest. Conversely, an analysis at the individual level describes a composite of within-groups and between-groups effects that is easy to misinterpret.

Ecological psychology

The context of human behavior is receiving increasing substantive attention in psychology.
Roger Barker devoted a career to the study of behavioral settings, adopting the Lewinian position that the situation into which the individual moves does as much to shape behavior as the personality of the individual. Barker's interest was in the microecology. Bronfenbrenner (1974, 1975, 1976) is less concerned with immediate situations and more concerned with the totality of the individual's environment. Though each individual's cultural setting may be unique, from Bronfenbrenner's point of view there are statistical similarities in the experiences of individuals living in the same neighborhood or participating in the same community culture. Neighborhood data are to be considered in the same light as classroom data, and subjected to separate between-neighborhood and within-neighborhood analyses.

Bronfenbrenner suggests that the effect of an experimental intervention (e.g., a program for disadvantaged children) is likely to be small unless it radically changes the ecology of its subjects. Moreover, he questions the appropriateness of a strictly individual psychology; insofar as the subjects are part of the same interactive community, the community and its members may constitute a single "subject". That is to say, a treatment may have a substantial effect in one community and a negligible effect or the opposite effect in another.
Perhaps the contrast depends upon characteristics that could have been identified at the outset of the intervention, or perhaps on fortuitous

occurrences. Bronfenbrenner and C. R. Henderson (personal communication) have embarked on a program of disentangling between-groups and within-groups components of variance that apparently is probing more into technical, statistical issues than I have.

Evaluative studies and school-effect studies

The importance of group effects is slowly becoming recognized in the literature on educational evaluation. As long ago as 1967, the Wiley-Bloom-Glaser debate took place at a conference on evaluation.

In the same year Bock and Wiley (1967) argued that the best design for a comparative educational experiment is often to assign classrooms -- not pupils -- to treatments, at random within schools. In data they studied, the component for differences among pupils usually accounted for 20-30 per cent of the sampling variance of the outcome mean within a treatment, with pupils regarded as random. The remaining 70-80 per cent of the sampling variance arose from schools and from classrooms within schools. In their data, classrooms accounted for far more variance in arithmetic fundamentals than schools, whereas schools (neighborhoods?) accounted for more variance in reading.

The Bock-Wiley paper is one of a series of "school-effect" studies that ask how much variance in achievement, aspiration, or career level is "attributable" to school differences. Another kind of school-effect study relates specific characteristics of the school (verbal ability of teachers, say, or mean sense of efficacy in the student body) to outcomes, as in the Coleman report.

Werts (1968), among many others, reacted critically to the Coleman report. Coleman's analytic device was to partial the school means on family background and similar student characteristics out of the final achievement test scores of individuals. The residual school effect (percentage of variance in individual achievement accounted for) was then interpreted as an index of the impact of the school. The Coleman report left the impression that excellence of facilities and other supposedly valuable features of the school program had little or no correlation with competence of graduates. Good students tend to be found in schools that have good facilities, hence the variance due to treatment overlaps the contribution of student ability. Partialling out student differences at the first step arbitrarily assigns the overlapping variance to student characteristics and not to treatment. Werts advocated a partitioning due to McNemar which evaluates the unique contributions of student characteristics and school characteristics and leaves their overlap as a third fraction of the predictable variance in outcome. The distinction between the individual and group levels of analysis was touched on in several of the papers, but became a focus of attention only recently. Coleman himself (1975) has now acknowledged the validity of the objection to partialling out family variables in estimating school effects, but says that critics misperceived the purpose of his 1966 analysis.

Luecke and McGinn (1975) contrasted Coleman's method of analysis with another adopted by Project Talent. Among other conclusions, they report that Coleman's method systematically overestimates the effect on achievement of family background (vis-à-vis effects of teacher and school quality). Their procedure is to simulate the generation of student achievement over five stages (years of schooling) by setting up a causal model and specifying parameters of the causal variables including the correlations that represent causal links. Far more is at issue in their paper than the level of aggregation. They are concerned with "dynamic" effects in a long-continued process, and with the fact that students change classes and schools. Scoring the quality of the student's own teacher instead of using the aggregate quality of teachers in his school also modifies conclusions. I find myself discontented with the Luecke-McGinn presentation, particularly in their use of certain zero-order correlations (e.g., family with first-year achievement) as the standard index of influence against which other analyses are judged. The study does illustrate the potential of simulations for forcing social scientists to recognize the consequences of their analytic decisions. Simulations may have important uses in later work on aggregation *per se*.

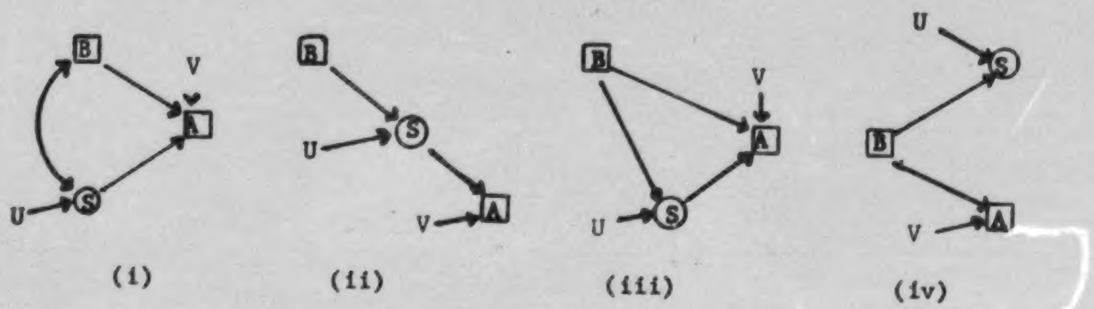


Figure 2.3. Alternative causal models for the relation of school quality to achievement

Duncan (e.g., 1970) has been especially insistent that any plan of analysis rests on a particular causal model. Pedhazur (1975) has recently applied similar thinking to the effect of school-level variables on achievement. Figure 2.3 is a modified version of a figure on his page 247. I use circles for global school-level variables and squares for individual-level variables (which may be aggregated). Otherwise, the diagrams follow the conventions of structural models. B represents background factors such as parental income; S represents a school-quality variable such as per-pupil expenditure or quality of teachers; and A represents student achievement. U and V are unspecified causes or sources of error. The B-to-S relation in (i) implies that neither variable causes the other; the relation must arise from some prior cause. In each other diagram, B contributes to S, perhaps via parents' willingness to vote more money for the schools. Achievement is said to depend directly on S in (ii), and directly on B in (iv); in (i) and (iii) the effect of B and S is joint. Pedhazur has much to say about alternative ways of partitioning variance and of testing the adequacy of the several models, but he does not emphasize units as such. (Later in his article, he does identify the main sources in the units-of-analysis literature but adds little of his own.)

If (iv) is the model, analysis is wholly at the individual level with B as the sole predictor. Adding S to the regression equation (with the same value for every student in the school) would falsely reduce the apparent contribution of B to A, and would imply that A depends on S. If this model does apply, the partial covariance $s_{AS \cdot B}$ will depart from zero only by chance.

If (ii) is the accepted model, analysis can appropriately be carried out at the group (school) level, with S as the sole predictor of the aggregate A. There would be no reason to supplement this with an analysis of the B-to-A relation within groups if one believes the model's assertion that the partial covariance $\sigma_{BA \cdot S}$ is zero.

If (i) is the accepted model, partitioning of the variance is open to the ambiguities discussed by Werts. The analysis can be done at the group level, with the aggregates B and A entered along with S. A supplementary individual-within-groups analysis of the B-to-A relation can be made.

Model (iii) is analyzed much as (i) is, but the interpretation can be less equivocal. The predictable variance of the aggregate A is allocated to two sources: B-independent-of-S, and S. The S portion includes some indirect or joint influence of B, but this is not seen as an influence "of B on A"; it is an influence mediated by school quality. Here again, the model makes it reasonable to consider a B-to-A-within-S effect at the individual level.

Most educational evaluations have been analyzed with individuals as the unit of analysis, even when using aggregate variables as in the Luecke-McGinn simulation. An example among studies of real data is the report on Head Start Planned Variation (Featherstone, 1973) in which the number of individual children receiving each treatment variation was taken as the sample size for it. The Performance Contracting experiment, on the other hand, used the average for a single school as a data point. Analysis at the classroom level is a third possibility -- as in the Maier-Jacobs and HPP studies.

In principle, the evaluator could recognize hierarchical nesting but apparently no evaluation report has done this.

Pupils are nested within classrooms, classrooms are nested within schools, and schools are nested within districts. It is reasonable to expect that an innovation mandated from the top of the hierarchy -- let us say, by the State Department of Education -- will not trickle down uniformly to the pupils. Possibly districts or communities will have a strong mediating influence, in causing the innovation to work or in sabotaging it; desegregation again comes to mind as an example. Innovations also succeed or fail at the level of the school, in the sense that strong school leadership can produce results whereas passive compliance wipes out the effect. Within the school, individual teachers conform or fail to conform to the treatment specifications, and add variation by the manner in which they carry out the treatment. And finally, of course, one expects individual differences within a classroom.

If one confines attention to the mean outcome in a treatment, nothing can be learned from a hierarchical breakdown; properly weighted, a mean is a mean is a mean. It is in the variances and regression coefficients at the several levels that differences could appear.

Extrapolation in interpretation

To conceive of an interaction "at the individual level" when treatments have been applied to individuals within groups is to engage in treacherous extrapolation. Data collected on groups are being explained in terms of processes within the individual. Such an interpretation is conventional in educational research and not unheard-of in general psychology, but it is highly questionable.

A treatment may be significantly altered by the very fact that it is administered to the subject when he is in company with other subjects. In the example above, the inductive procedure for teaching history would be radically altered by applying it to students individually, as that would allow no way to retain the important feature of group discussion.

The "operational definition" of the treatment consists essentially of a set of instructions directing the acts of the experimenter or teacher. When this identical operation is shifted from a group context to an individual context, the treatment is likely to be significantly altered. The teacher's reprimand, or instruction to "Pay attention to . . .", is a different stimulus when addressed to the group in general than when addressed to the student in isolation. Thanks to social facilitation, doing a page of arithmetic drill alongside one's classmates is not the same task as doing it alone. Thus the operational definition has to specify individual administration or group administration -- more than that, it has to specify the basis for constituting groups.

As is well known, evidence collected from applying one operationally defined treatment may indicate little or nothing about what will happen when a treatment with another operational definition is administered. Sometimes the change in the operation (here, the change made by altering the context) makes a large difference, and sometimes it makes none. This has to be tested directly; experiments with one operation do not give direct evidence on another. One may reason indirectly *if* he has established a strong presumption that the change in operation never matters. Bergmann (in Frank, 1961, p. 53) uses the example of the location of the apparatus in an experimental room. The presumption that a shift in location makes no difference is so strong, Bergmann says, that we are willing to ignore a shift in that aspect of the specification. Bergmann is right in the abstract, but his example is telling in a way he did not intend.

Gerald Holton tells me of the experience of Fermi and his group when they first attempted to bombard the nucleus with neutrons, in an attempt to create artificial radioactivity. They got negative results when the apparatus was set up on one bench in the laboratory, and got success on another bench. The critical difference was a marble surface on the first bench. Neutrons rebounding from the surface had no more effect than those directly fired at the target nucleus. The second table, with a wooden surface, slowed the neutrons while scattering them, and it was these rebounding slow neutrons that produced the effect Fermi was seeking. Holton tells a similar story of Rutherford, discovering thorium. Rutherford's electroscope discharged when far from the open door of his laboratory because of radioactive gases in the air. When he had collected data with the apparatus near the door there was no discharge; the gases were swept away by currents near the open door. A presumption that a certain shift in operation has no effect, then, is a presumption made at considerable risk.

Group contexts surely affect human behavior at times. Hence evidence collected by observing individuals behaving in groups is not a dependable indication of what will happen in an individual experiment. Nor can evidence obtained in groups composed in one manner indicate what will happen when the groups are formed by a different procedure, unless a strong theory about the character of the context effects has already been worked out.

The group-level and within-group effects are observed in a sample of classes. These classes can be regarded as drawn from a population of classes formed by a certain process. The results can be generalized to that population of classes, i.e., to classes formed by the same process from a similar pool of persons. The inference is of a type commonplace in statistics. To make an inference to classes formed by some other process or rule is just as much a leap in the dark as it would be to extrapolate from the treatment observed to some variant of it.

The experimenter may or may not know what process formed the classes he observes. If he formed them by randomly grouping members of a student body or by another formal assignment rule, he can generate similar groups by that same process so long as the population of students is unchanged from year to year. If the groups were formed by the existing community and within-school processes, his findings will apply so long as the school population is stable and those processes continue to control class membership. He cannot assume that the findings will apply if a new grouping procedure is installed following the experiment. Altering nothing but the size of the instructional groups might be enough to change the relation of interest.

This line of argument can lead in two directions. (1) To be conservative, the person conducting an experiment in intact classes will limit his conclusion to classes formed in the same manner. He should specify the process that formed the classes or the characteristics of the classes, in such a way that others making use of his research can judge whether their classes resemble his in composition. This extends the usual recommendation regarding description of a subject population. In classroom experiments, the class is the subject and the characteristics of the subject classes should be brought into the open. (2) A liberalizing step is to regard the assembly rules as treatment dimensions. In the course of a long program of work, particularly work oriented toward theory, an investigator varies the specifications for an experimental treatment. The successive variants form a collection that can be described by parameters, and the varying effects can be described as a function of the parameters. When this process is well advanced, the investigator can make reasonable predictions about treatments that have not been roadtested. Just as a collection of manipulated treatments has parameters, a population of classes has parameters. Applying different sorting processes in successive experiments would build up some theory. One might be able then to make limited extrapolations to classes formed in ways other than those directly observed.

#

3. A mathematical model

The regression equation needed for distinguishing the effects of interest can be built up in steps. Confine attention to a single treatment, and identify persons p as members of groups c . The person has scores X_p and Y_p , which may for emphasis be written X_{p_c} and Y_{p_c} . The model could be set up in terms of ${}_X\mu_p$ and ${}_Y\mu_p$, the "true" or "universe scores" that would hypothetically be obtained by exhaustive measurement. In this section the model is in observed-score form. Universe scores will be taken up in Section 6. The class mean μ_{Y_c} is the mean over the fixed class of Y_p $(p \varepsilon c)$; likewise for X . I should note also that, so long as questions of statistical inference are held in abeyance, the model applies to dichotomous variables as well as to continuous ones.

This section can be read as if all collectives have the same number of members. When the number is variable, the definition of any parameter involves a weighting decision. See Section 4.

Definition of components

The Y score may be divided into general-level, between-group, and within-group components in the usual manner:[1]

$$(3.1) \qquad Y_{p_c} = \mu_Y + (\mu_{Y_c} - \mu_Y) + (Y_{p_c} - \mu_{Y_c}) \; .$$

The between-groups component divides into a part predicted by the group mean on X and a residual.

$$(3.2) \qquad (\mu_{Y_c} - \mu_Y) = \beta_b \, (\mu_{X_c} - \mu_X) + \gamma_c \; .$$

It is to be noted that the same regression coefficient serves to predict Y_{p_c} from μ_{X_c} .

The within-groups effect is decomposed in two stages. Write β_w for the common within-groups regression coefficient that best accounts for the sum of squares within groups. Then

(3.3) $$(Y_{p_c} - \mu_{Y_c}) = \beta_w (X_{p_c} - \mu_{X_c}) + \delta_{p_c}$$

But within a particular group the regression slope β_c need not equal β_w, which leads to the further decomposition:

l.c.
delta
epsilon

(3.4) $$\delta_{p_c} = (\beta_c - \beta_w)(X_{p_c} - \mu_{X_c}) + \varepsilon_{p_c}$$

Putting the equations together gives this series of components:

(3.5) $$\begin{aligned} Y_{p_c} - \mu_Y &= \beta_b (\mu_{X_c} - \mu_X) && \text{Between, predicted} \\ &+ \gamma_c && \text{Group residual} \\ &+ \beta_w (X_{p_c} - \mu_{X_c}) && \text{Common within, predicted} \\ &+ (\beta_c - \beta_w)(X_{p_c} - \mu_{X_c}) && \text{Specific within, predicted} \\ &+ \varepsilon_{p_c} && \text{Person residual} \end{aligned}$$

The overall slope β_t considered by those who analyze at the individual level is a composite of β_b and β_w. As shown by Duncan, Cuzzort, and Duncan (1961, p. 66):

$$\beta_t = \beta_w + \eta^2_X(\beta_b - \beta_w) \quad \text{or} \quad \eta^2_X\beta_b + (1 - \eta^2)\beta_w$$

where η^2_X is the intraclass correlation of X [equal to $\sigma^2(\mu_{X_c})/\sigma^2(X_{p_c} - \mu_X)$]. 1a

Coupled with the argument on pp. 4 .3-4, this formula has an important implication for those who try to interpret β_t in typical educational studies. β_t is a weighted average. In studies with a modest number of groups, β_b is badly estimated, though β_w may be well estimated. The larger the value of η^2, the more the error in $\hat{\beta}_b$ makes for errors in $\hat{\beta}_t$. (The investigator usually has the illusion that numerous cases entered into the latter calculation.)

Putting μ_Y on the left side of (3.5), I leave it unanalyzed.

In a two-treatment study, μ_Y includes the treatment effect as well as the general mean. In the two-treatment study of the usual type, groups are nested within treatments and the treatment constitutes a third level in the hierarchy.

The model (3.5) could be defined with $E\beta_c$ replacing β_w. In general, this change decreases the "common within, predicted" variance and increases the specific-within variance. The definition in terms of β_w is more conventional, often being associated with an assumption that group membership is random and the Y, X distributions within classes homogeneous.

The distinction has little importance for descriptive statistics. For some sets of data we have calculated b_w and $\overline{b}_c$ and found that the two differed negligibly. The expected value

$$E\beta_c = \underset{c}{E}\,[\sum_{\substack{p\\p\in c}} (X_p - \mu_{X_c})(Y_p - \mu_{Y_c}) / \sum_{\substack{p\\p\in c}} (X_p - \mu_{X_c})^2]$$

does not generally equal

$$\beta_w = [\underset{p}{E}\,(X_p - \mu_{X_c})(Y_p - \mu_{Y_c})]/\underset{p}{E}\,(X_p - \mu_{X_c})^2 \ .$$

If more than one X is available, for every subject, weighted composites will account for more variance, between and within groups, than the single X . The model can be extended by introducing, for example, a β_{b1} , β_{b2} , etc. It is comparatively difficult to think about the multivariate case, however; the best composite predictor between groups may differ from the best within-groups composite, and each predictor group may have its own best composite. The usual intuitive understanding of multivariate relationships is confounded by the fact that predictors whose between-groups correlation equals zero may have a nonzero within-groups correlation, or vice versa. Consequently, any geometric analogy (e.g., reference to "dimensions") is likely to go astray. I shall return to the two-predictor problem.

A simple structural model will perhaps be helpful (Figure 3.1). The grouping rule determines the division of X between the class mean and

at 3.3

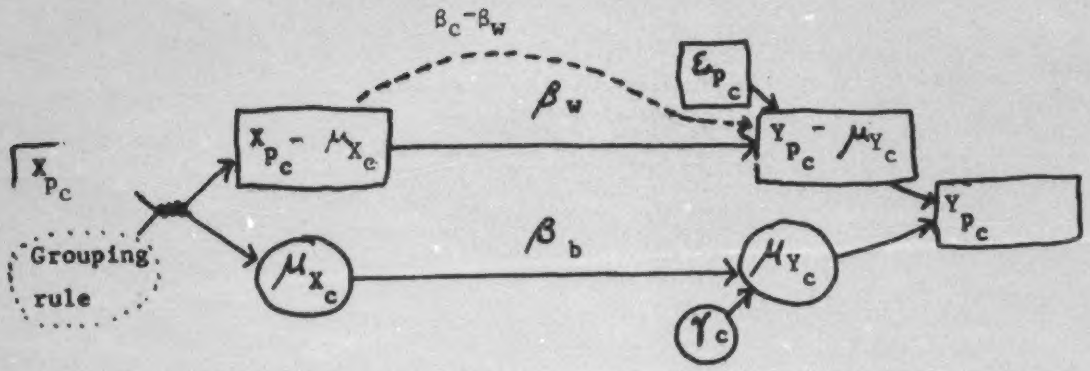

$\beta_c - \beta_w$
β_w
ε_{p_c}
X_{p_c}
$X_{p_c} - \mu_{X_c}$
$Y_{p_c} - \mu_{Y_c}$
Y_{p_c}
Grouping rule
μ_{X_c}
β_b
μ_{Y_c}
γ_c

Figure 3.1. Structural model for hierarchical analysis.

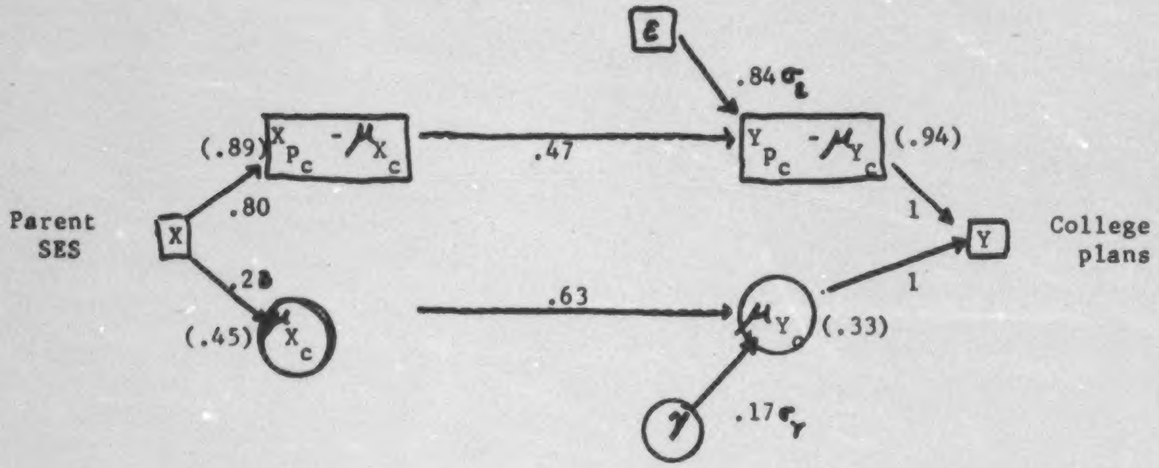

ε
$.84\sigma_\varepsilon$
(.89)
$X_{p_c} - \mu_{X_c}$
.47
$Y_{p_c} - \mu_{Y_c}$
(.94)
Parent SES
X
.80
1
Y
College plans
.28
1
.63
(.45)
X_c
μ_{Y_c}
(.33)
γ
$.17\sigma_\gamma$

Figure 3.2. Structural regression based on data of Campbell and Alexander Numerals show unstandardized regression coefficients; numerals in parentheses show standard deviations for selected components.

the deviation score. These two are uncorrelated. Analysis may then proceed separately in the upper and lower tracks. The conventional "individual-level" analysis can be thought of as proceeding in the same manner, save that β_b and β_w are constrained to have the same value.

The concrete example in Figure 3.2 is derived from a figure of Duncan *et al*., 1972, p. 193; c here symbolizes a school, not a class. The original data were supplied by E. Q. Campbell and C. N. Alexander, Jr. (N = 1137.) Duncan *et al*. give correlations; I assume $s_X = s_Y = 1$ to get regression coefficients. Also, I assume linearity (i.e., that $\eta_X = r_{X_p} \mu_{X_c}$). The intraclass correlations for X and Y are 0.20 and 0.11, respectively, indicating that more of the variance lies between schools in the independent variable than in the dependent variable. This on its face suggests that schools do not cause divergence. Such an inference would be stronger if it were believed that SES is a sufficient specification of the precursors of educational aspirations established at the time students entered school (including any preliminary statement of aspirations). The regression coefficient, however, is higher between schools than within, which on its face argues *for* a tendency of high SES schools to cause aspirations to rise. (For more on this kind of reasoning, see p. 3.18.) The regression coefficients here are consistent with the difference in correlations noted by Duncan *et al*. (0.86 between and 0.45 within); this would not always be the case.

Interpretation of components. To give a further sense of the implications of the five basic components, Figure 3.3 displays regression lines like those of Figure 1.1. Regression lines for two groups are represented in Panels (i)-(iii). (The lengths of the lines have no significance.) With two groups, the group means fall on the between-groups regression line and γ_c is

at 3.4

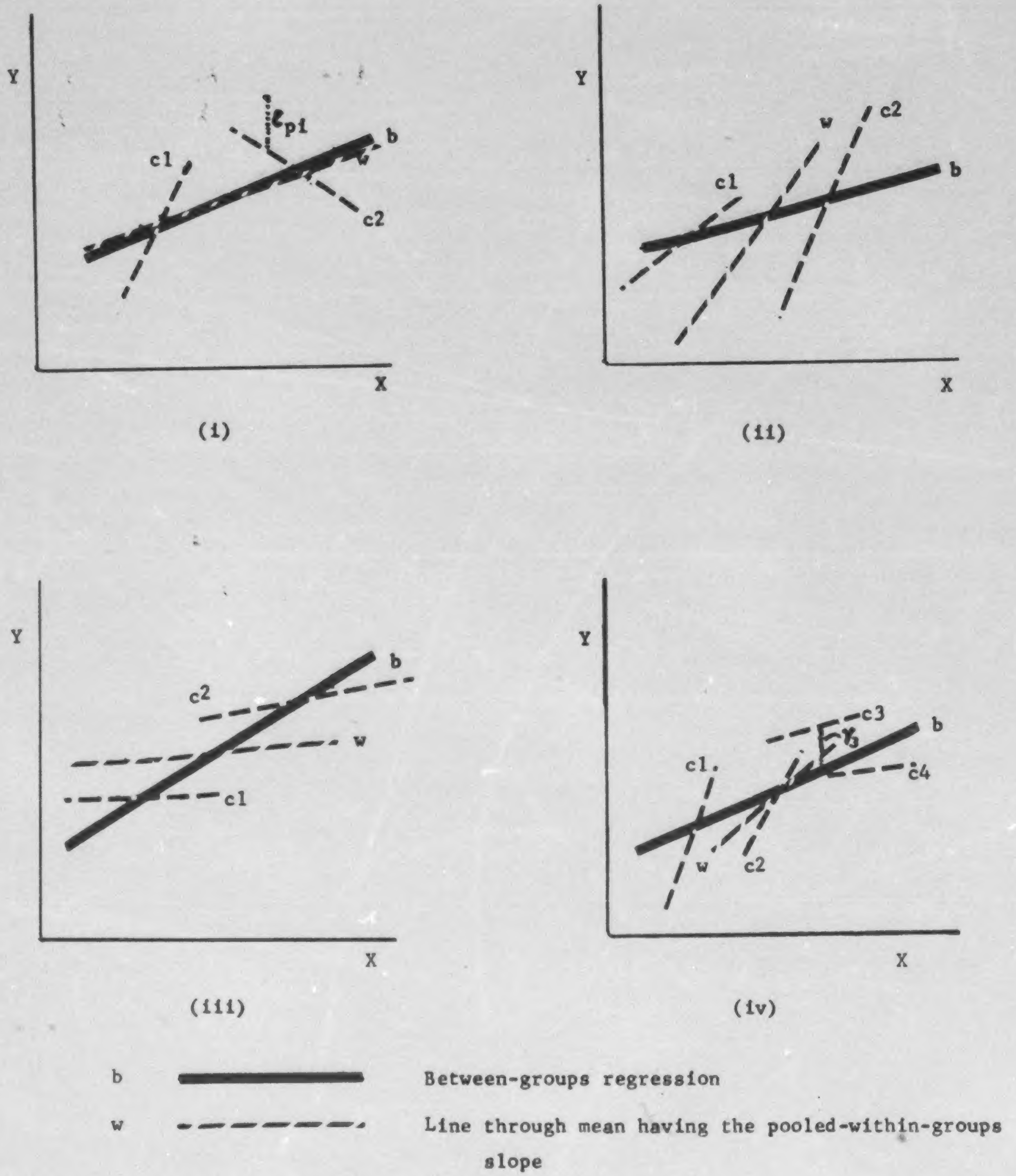

Y
X
c1
c2
b
w
ℓpi
(i)
Y
X
c1
c2
w
b
(ii)
Y
X
c2
c1
w
b
(iii)
Y
X
c1
c2
c3
c4
w
b
γ3
(iv)
b — Between-groups regression
w — Line through mean having the pooled-within-groups slope
c — Specific within-group regression

Figure 3.3. Possible relations among β_b, β_w, and β_c.

necessarily 0. In Panel (iv), four groups are shown. One of the γ components is labelled. In Panel (i), the ε component for a single member of Group 2 is identified.

Regarding the residuals: The residual ε_{p_c} includes any effects of individual characteristics p brought to the experiment that were not fully represented in X, and errors that caused his observed Y to differ from his universe score on Y. (As I have stated the model, the average error over members of his class is added into μ_{Y_c} and removed from ε .) Furthermore, it includes any effect, unpredictable from his X score and not common to other group members, that influenced his final level of accomplishment -- an illness, for example.

The residual γ_c can be thought of as an adjusted "group effect." Where the group is a class, γ_c includes the "main effect" of the teacher, plus the effects of variations in the delivery of the treatment to this class, plus unpredicted effects that have a net influence on the mean (uncommon enthusiasm, an epidemic, etc.), and the average error of measurement in the Y_p . The individual-level errors need not have an average close to zero.

The between-groups effect described by . β_b reflects any consistent tendency of higher-X groups to do better than others (or worse) on the outcome measure. An example already mentioned is the possibility that teachers cover more ground in abler classes.

The common-within effect reflects the tendency for students above the group average to outperform (or underperform) the rest of the group. The regression coefficient is derived from data on all groups combined; it need not fit any one group. The educator's usual interpretation of the effect of aptitude on outcome is that students above the grand mean do better, regardless of their classmates. That amounts to a prediction that β_w will be positive. If students are assembled into groups on the basis of X information alone, and working in the groups makes no difference, β_b will equal β_w, as suggested in Panel (i). But the inference cannot be reversed. The fact that $\beta_b = \beta_w$ does not identify the causal situation (see p. 3.17).

The specific within-group regressions are likely to vary, but it is hard to know whether to take this variation seriously. The regression coefficients will differ by chance even when the processes operating in the classes are basically the same. Second, insofar as the selection factors operating to form the several groups differ, the slopes will be affected. Third, and most interesting, are the possible differences in causal processes. Slope differences might come about, for example, if one teacher distributes attention to high-X and low-X students in a different proportion than the next does, or if some teachers set up a strong competition that encourages the able and discourages the others.

The configuration in Panel (ii) suggests that instruction in high-X groups differs little in average effect from that in low-X groups. Within groups, however, the student's X level makes a substantial difference. Apparently, the treatment has set up a scarcity economy within the classroom, so that the comparatively able students snatch up the educationally useful experience at the expense of the comparatively weak students. If these are the results, a student near the average of the overall X distribution is much better off in a class that, as a class, is below average on X. The student with high X would accomplish far more in a wide-range class, but such a class can be constituted only by bringing in low-ability students, who are sacrificed in his interest. The students with low X scores accomplish more if placed in a homogeneous low-X group.

A grouping policy is derived only by extrapolation, however. Would a strictly homogeneous group of students with uniform low scores on X fall on the between-groups regression line found in this experiment on heterogeneous groups?

Panel (iii) shows β_w less than β_b .

[Here, the student's final outcome depends a great deal on the level of the class, and little on his comparative standing within the aptitude distribution of the class. Perhaps such a configuration describes what would be found if the graduates of a prestigious medical or business school and a run-of-the-nation school were assessed. The highly selective school probably offers a more intensive program of training. Once the program is adapted to the level of the group, it may be that factors other than aptitude X account for differences in success within either group. Thus X might have been a valid predictor of success before the school programs diverged, and not after.

The slopes representing β_c have various configurations. In Panel (i), $\beta_{c2} - \beta_w$ is negative. Something happening in this particular class negated the advantage abler students have within typical classes, represented by the slope β_w .

[Tedious instruction might have this effect.

It will be useful to recapitulate much that has been said by reproducing Figure 3.1 in the form of Figure 3.4, with labels attached to the causal connections. The labels are illustrative and not exhaustive; chance effects also enter the residuals. I have not separated the specific and common-within effects here.

at 3.7

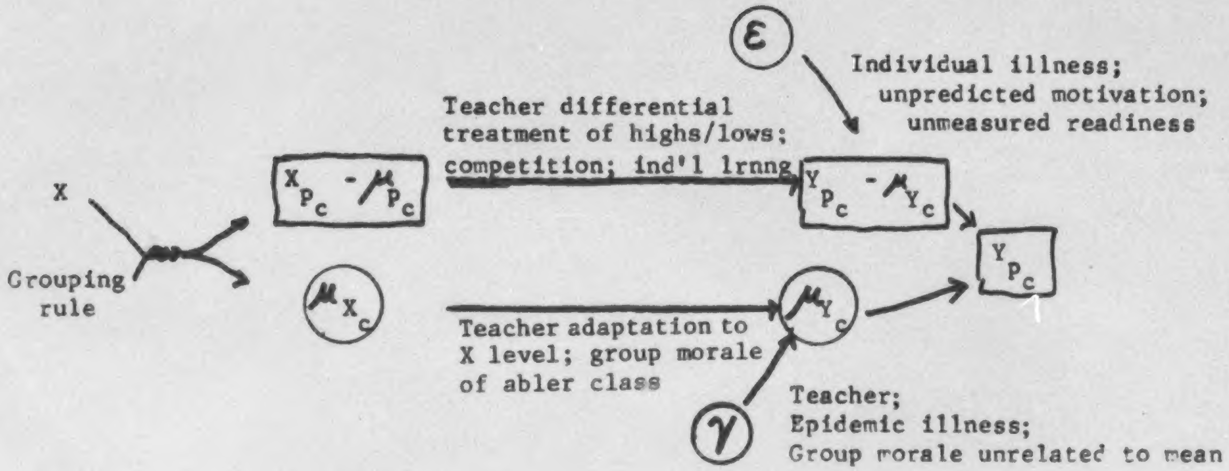

ε
Individual illness;
unpredicted motivation;
unmeasured readiness
Teacher differential
treatment of highs/lows;
competition; ind'l lrnng
X
Grouping rule
$X_{P_c} - \mu_{P_c}$
μ_{X_c}
$Y_{P_c} - \mu_{Y_c}$
μ_{Y_c}
Y_{P_c}
Teacher adaptation to
X level; group morale
of abler class
γ
Teacher;
Epidemic illness;
Group morale unrelated to mean

Figure 3.4. Interpretation of causal arrows in Figure 3.1

What is easily overlooked is that Grouping Rule is a causal factor. Change the rule, and all the regression coefficients would change. This perhaps says no more than that any coefficient applies only to a certain population of groups (Section 4); but the present formulation, along with what is to come at page 3.11, emphasizes that the rule for assembling groups is often a manipulable, causal variable. The intraclass coefficient for X is the regression coefficient that goes with the causal arrow leading to μ_{X_c}, and $1 - \eta_X^2$ is the coefficient leading to the deviation score.

Partitioning variance

The model leads directly to a partitioning of variance. The overall variance of Y divides into between-groups and within-groups portions, and [each of these subdivides. If one assigns to each individual three predictor scores -- X_{p_c}, μ_{X_c}, and c, where the last is a string of coded dummy variables -- a stepwise regression analysis will decompose the variance of Y. Predictors enter in the order shown at left, and the change in the mean square for regression at each step estimates a variance:

μ_{X_c}	Variance attributable to the between-groups regression effect	$\beta_b^2\,\underset{\smile}{\sigma}^2(\mu_{X_c})$
c	Remainder of between-groups variance	$\underset{\smile}{\sigma}^2(\gamma_c)$
X_{p_c}	Variance accounted for by common within-groups regression	$\beta_w^2\,\underset{\smile}{\sigma}^2(X_{p_c} - \mu_{X_c})$
$X_{p_c} \times c$	Additional variance accounted for by specific within-group regression (Group ×Aptitude interaction)	$\underset{c}{\text{Av.}}(\beta_c^2 - \beta_w^2)[\sigma^2(X_{p_c} - \mu_{X_c}) \mid c]$
(Remainder)	Unpredicted individual variance	$\sigma^2(\varepsilon_{p_c})$

The variances indicate the potency of the five components to produce differences in Y. The variances for $X_{pc} - \mu_{X_c}$ and ε_{p_c} are defined over all groups pooled.

Choices made in forming the model

In setting up any model one chooses between alternatives.

Direction of decomposition. I have chosen to partition group effects out of the Y scores before evaluating within-group relationships. The stepwise order shown above is just one of the possible orders(discussed e.g. by Werts, Duncan, or Pedhazur). The variance that might be attributed to the overlap of group and individual characteristics is assigned here to the between-groups component. Insofar as this is an arbitrary choice, not justified by a causal model, it generates ambiguity in interpreting $\beta_b - \beta_w$.

It would have been possible to set up the model in just the opposite way, fitting a regression line to individual scores without regard to groups and then asking if group membership accounts for additional variance and covariance (see p. 1.17b, 1.22). This is a substantive decision. A considerable amount of variance in instructional practice occurs at the group level. It seems to me that individual differences are best studied by comparing persons treated under the same circumstances. Members of the group are, in a sense, in the same circumstances. In multilevel hierarchies, effects could be removed in many orders; only a theory about particular variables justifies one model rather than another. Thus it might be argued, with respect to change in interracial attitudes after desegregation, that the school is a more critical unit than the class or the district. It might equally be argued that in improving the self-concept of individual children the class is a more potent source of variance than the school. Some other variable (truancy?) is perhaps more associated with the individual and his home, and less with the unit of instruction. Then, it would make sense to frame the model with the individual as primary. Such a model might start with an analysis of the pooled data from all classes, and then look at groups with unexpectedly high and low rates of truancy.

Nonlinearity. Nonlinear terms could be added to the model. After extracting the common within-groups regression, one can reasonably fit a coefficient to terms of the form $\beta_c(X_{p_c} - \mu_{X_c})(\mu_{X_c})$; the contribution of the specific-within-group regression is reduced accordingly. This added member allows for the possibility that within-group slopes are linearly related to the class mean on X (as in panel iv of Figure 1.3). Whether it will be profitable to make this separation is to be judged in the light of one's prior beliefs about the phenomenon under investigation.

It leaps ahead of the story to consider predictors other than X. Any effect of global properties of groups on within-groups relationships must come via a product term. A group property G_c is necessarily uncorrelated with $Y_{p_c} - \mu_{Y_c}$, but $G_c \times (X_{p_c} - \mu_{X_c})$ may correlate with $Y_{p_c} - \mu_{Y_c}$.

Nonlinearity could also be introduced via quadratic terms [in $\mu^2_{X_c}$, $(\mu_{X_c})(\mu_{X',c})$, $(X_{p_c} - \mu_{X_c})^2$, etc.]. In fact, one of the most striking findings of distinctive within-class regressions is that of Majasan (see Cronbach, 1975). Majasan predicted that measured achievement in a college psychology class would have a parabolic regression on the students' BQ scores. The BQ score reflected acceptance of behavioristic (vs. humanistic) statements. Majasan had a BQ score for each instructor and he predicted that (with aptitude held constant) the parabolic regression would have its peak where the student BQ matched the instructor's. He was able to confirm this prediction in 10 out of 11 classes, the exception being a class where no measured-achievement criterion was available. (Majasan could not investigate the between-class regression because the course examination varied with the class.)

There is a lively danger that regression techniques will dramatize relationships that arose by chance; and making hypotheses complex adds to the risk. Nonlinearities may reasonably be explored, but unless there is a rationale for predicting nonlinearity, little credence can be given a nonlinear relationship the first time it turns up.

Effects of aggregating data

It will be necessary next to examine the relation among β_b, β_w, and β_t. The three are linked by the intraclass correlation η^2 (p. 3.2). My ideas on this subject have been formed over years of discussion with Leigh Burstein, whose dissertation (1975) on the bias problem has in turn been influenced by his work with Hannan (Hannan & Burstein, 1974). My formulation is structured differently from Burstein's in important particulars, but the formulas to be presented for β_b in Figures 3.5 and 3.7 are consistent with his.

The traditional problem of "assessing the bias" due to analyzing at the group level when β_t is wanted deserves little of our attention -- we rarely want β_t. We do want to compare β_b and β_w.

The development that follows (and the highly general development that ends the section) lays out some algebraic tautologies. It does not depend in any way on substantive considerations, and would hold true for data generated by any causal model whatsoever. The analysis nonetheless fulfills an important function, in showing how numbers that are sometimes given a substantive interpretation can be generated by the aggregation rule.

The argument is most directly understandable when two individual characteristics that exist simultaneously are to be related to each other, for example the potato and wheat yields of Yule and Kendall, or the ethnic and religious identifications discussed by Duncan *et al*. This is *post hoc* grouping; the effects have already been developed, perhaps in group settings that have no relation to the groups now being composed for purposes of analysis. Those data might be grouped in a number of arbitrary ways; the joint distribution overall has been established. For the Y-on-X regression, how does the within-groups or the between-groups regression depart from the overall β_t ? The counterpart question can be asked about the X-on-Y regression. The answers will depend on how the bases on which groups are differentiated relate to X and Y, or, what is equivalent, to X and the partial variate $Y \cdot X = Y - \beta_t X$.

Special case with linear assumption. To develop a comparatively simple argument, I introduce a discriminant function W , and assume that X, Y, and W have a multivariate normal distribution. We may consider that groups are formed by dividing W into regions and assigning persons in the same region to the same group for purposes of analysis (not necessarily for treatment). It follows that the means of X, Y, and W are perfectly correlated. (It is this condition on which the immediately following argument depends. It could be satisfied when the assumption of trivariate normality is not. But such a strong assumption will serve for the moment.) In the rest of the argument I shall use Z and not W; Z is simply the group mean on W, and every member of the group has the same Z score.

Instead of working with correlated variables I substitute orthogonal variables I, II, and III; these are perfectly correlated, respectively, with X, $Y \cdot X$, and $Z \cdot Y, X$. Each of these components has a zero mean and a unit s.d. over individuals in the population. The relations to be developed in this population would be found for samples, if sample statistics were used to define I, II, and III. Variables I, II, and III can be thought as coordinates in a three-space; I and II are orthogonal coordinates for the X, Y plane.

Figure 3.5 shows how the standard scores on X, Y, and Z may be described in terms of component loadings. (To use standard scores simplifies the argument without loss of generality.) The symbol A is used for the correlation of X and Y, overall, and CB for the correlation of X and Z. This notation is used because B proves to be a key parameter; all Z that project into the same line of the X, Y plane have the same B. Since dimensions I and II carry the information in X and Y, and C^2 is the proportion of variance in Z that I and II account for, $C = R_{Z \cdot X,Y}$. As a convention, C takes a positive sign.

If $C = 1$, Z lies in the X, Y plane and therefore must be a continuous variable. This can occur only if the grouping procedure sliced the W scale into infinitely thin slices, hence it is a hypothetical limiting case.[2]

As Figure 3.5 shows, the covariances within any group (i.e., among persons with Z constant) can be described by a partial-covariance formula. Then simple subtraction produces the between-group covariances in the population. No matter what the value of C the between-groups regression coefficient is the same. The relation of β_t to β_b, then, depends on B and not on A or C.*

This formulation applies to the population. As with $C = 0$, $C = 1$ is a hypothetical limiting case. A finite collection of individuals can be regarded as a population, hence the formulas can apply to them. If the grouping procedure is strictly random, however, the correlation of X with W and that of Y with W will not be precisely zero. Consequently, C will not reach 1. With purely random post hoc assignment, any one group is a random sample of the total collection of individuals and its within-group regression coefficient is an unbiased estimate of β_{YX}. Consequently, with purely random grouping subsequent to treatment of individuals,

at 3.14

---

*In terms of A, η_X, and ϕ, the formulas for β_b and β_w remain the same, when the linear restriction is removed (p. 3.24).

	I	II	III
X	1		
Y	A	$+\sqrt{1-A^2}$	
Z	CB	$C\sqrt{1-B^2}$	$\sqrt{1-C^2}$

β_{YX} overall $=$ A

β_{YX} within groups $=$ $\beta_{YX \cdot Z}$

$$= A - \frac{BC^2\sqrt{1-A^2}\ \sqrt{1-B^2}}{1 - C^2B^2} \quad \text{or}$$

$$A - \frac{\eta_X^2}{1-\eta_X^2}\sqrt{1-A^2}\ \tan\phi$$

β_{YX} between groups $= A + \dfrac{\sqrt{1-A^2}\ \sqrt{1-B^2}}{B}$

or $A + \sqrt{1-A^2}\ \tan\phi$

$\tan\phi = \dfrac{\sqrt{1-B^2}}{B} = \dfrac{\eta_Y - \eta_X\rho_{XY}}{\eta_X\sqrt{1-\rho_{XY}^2}}$ See Note 2.

β_{YX} between groups $= \eta_Y/\eta_X$

Overall covariances

	X	Y	Z
X	1	A	CB
Y		1	$CAB+C\sqrt{1-A^2}\ \sqrt{1-B^2}$
Z			1

Within groups covariances (Z partialled)

	X	Y	Z
X	$1- C^2B^2$	$A - C^2AB^2 - BC^2\sqrt{1-A^2}\ \sqrt{1-B^2}$	
Y		$1-C^2[AB+\sqrt{1-A^2}\ \sqrt{1-B^2}]^2$	
Z	0	0	

Between groups covariances

	X	Y	Z
X	$C^2B^2 = \eta_X^2$	$C^2B[AB+\sqrt{1-A^2}\ \sqrt{1-B^2}]$	CB
Y		$C^2[AB+\sqrt{1-A^2}\ \sqrt{1-B^2}]^2 = \eta_Y^2$	$CAB + C\sqrt{1-A^2}\ \sqrt{1-B^2}$
Z			1

First part of Figure 3.5

at 3.14

Values of regression slopes when C = 1

B	-1	$-\sqrt{1-A^2}$	-A	0	A	1
Z coincides with	-X	X·Y		Y·X	Y	X
η^2_X	1	$1-A^2$	A^2	0	A^2	1
Overall slope	A	A	A	A	A	A
Between-groups slope[2]	A	0	2A-1/A	indet.	1/A	A
Within-groups slope[2]	indet.	1/A	2A	A	0	indet.

Figure 3.5. Regression coefficients as a function of parameters of a single grouping variable under a linear assumption

the values of β_w and β_b both approach β_t as the number of groups becomes larger.

How does the discriminant function affect β_b and β_w ? Figure 3.5 gives formulas in A, B, and C, and, more simply, in terms of ϕ . Let X* be the projection into the X, Y plane of the line along which the means $\overline{X}, \overline{Y}$ lie, and define ϕ as the angle between X* and X ; $\cos \phi = B$.

Figure 3.6 is for the case $\beta_t = 0.40$; a horizontal broken line represents β_t . One curve displays the value of β_b at each ϕ from -90° to 90°; this curve would be repeated in the range $90° < \phi < 270°$. A second function represents values of β_w with C = 1; this is the unrealistic case where W is divided into infinitesimal regions. The third function gives β_w when C = 0.8. As C declines toward zero the line for β_w comes closer to that for β_t .

Obviously, the relation of β_b and β_w depends on the relation of the grouping variable Z to X and Y. This is similar to the effect of "restriction of range" or "truncation" on test validity -- a problem well known in psychometrics. I interpret these results before offering a more general development.

Meaning of ϕ . Traditional writings on aggregation bias have thought of the grouping principle as one that could be arbitrarily established. Thus Feige and Watts (1972) were looking for a way to group banks into small sets so that the Federal Reserve System could report data for the sets without violating confidentiality, and yet the data would represent the microeconomic relations adequately. The formulas of this section apply to such a case, but we are also interested in applying them to groups that were formed by natural processes.

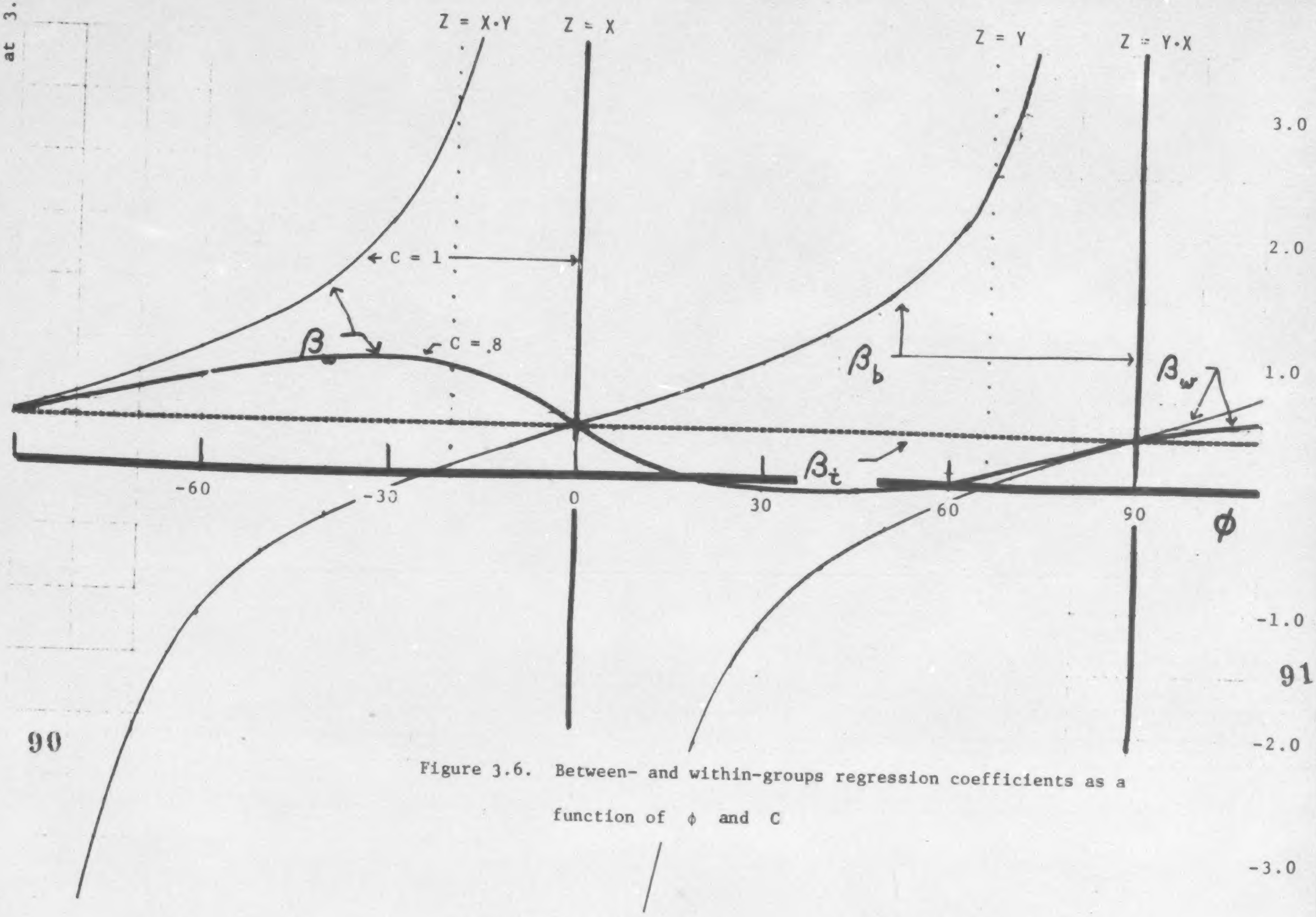


Figure 3.6. Between- and within-groups regression coefficients as a function of ϕ and C

β_t, which defines $Y \cdot X$, is determined from the overall distribution. In an experimental treatment, the values of Y reflect any predictable and unpredictable effects including context effects. The grouping variable may be related to $Y \cdot X$ for many reasons. The most startling paradox is that if groups are formed strictly on the basis of X there may nonetheless be a relation of Z to $Y \cdot X$, hence a non-zero ϕ. Suppose that grouping is based on X, and no other initial characteristic adds to the prediction of Y. But suppose there is a context effect, such that high-X groups have high Y means. If one partials X out of Y, using the overall regression coefficient, the high-X groups will have a positive residual $Y \cdot X$ and the low-X groups will have a negative residual. Then ϕ will be positive. (An inverse context effect, with high-X groups doing comparatively badly, will make ϕ negative.) A positive ϕ could also arise in the absence of any context effect if grouping takes into account some W that is correlated with individual values of $Y \cdot X$. An unlikely third possibility is that grouping is actually based in part on the outcomes of the treatment.

Comparing β_b and β_w. In the sociological literature there is a line of reasoning that runs like this: If β_b equals β_t, there are no context effects. Or, equivalently, if $\beta_b = \beta_w$, there are no context effects. Thus one can regress college plans, as in Figure 3.2, on SES within schools. If $\beta_b = \beta_w$, Duncan *et al*. (1972, p. 197) would dismiss the hypothesis that the school climate had an effect (direct or indirect) on aspirations. In such a case they suggest that the between-groups slope merely aggregates evidence of effects at the individual level.

Duncan *et al*. (p. 195) expand on their version of Figure 3.2 (in a manner that loses something in my compression and in translation to the regression form). Essentially, they substitute for the μ_X-to-μ_Y path a diamond-shaped configuration that makes room for intermediate variables $\tilde{Y}_c$ and Y^*_c , whose sum is μ_{Y_c} . $\tilde{Y}_c = b_w \mu_{X_c}$ and hence is a predicted school mean under the hypothesis $\beta_w = \beta_b$. In the population $\beta_w \mu_{X_c}$ would surely be the between-groups predictor *if classes had been formed at random and if no causal effects at the school level* entered into μ_{Y_c} or Y_p . The part of μ_{Y_c} that is left over, Y^*_c , is analogous to the adjusted mean in analysis of covariance.

For the Campbell-Alexander data the coefficient linking μ_{X_c} to $\tilde{Y}_c$ is set equal to b_w, i.e., 0.47; and the coefficient leading to Y^*_c is 0.63 - 0.47 = 0.16. The standard deviations of $\tilde{Y}_c$ and Y^*_c are 0.21 and 0.07, respectively. All this suggested to Duncan *et al*. that "composition effects" (demographic) are much stronger than "school effects" (group caused). This kind of inference does not appear to be justified.[3]

In the Campbell-Alexander data $\eta^2_X = 0.20$. Suppose that 0.50 is the regression coefficient of college plans on SES that would occur if each student were miraculously to grow through adolescence completely insensitive to any effect of the group around him, so far as his college plans are concerned. Under this no-school-effects hypothesis, if students were grouped nonrandomly with $\eta_X^2 = 0.20$, the values of β_b and β_w could be those of the Campbell-Alexander data: 0.63 and 0.47. What is required under the linear assumption of the model above is that

$\tan \phi = 0.15$, which implies that grouping depends much more on SES than on other correlates of educational aspirations. When grouping has no causal effect, a difference between β_b and β_w will occur unless $\phi = 0$. That is, a nonzero $\beta_b - \beta_w$ certainly implies a causal effect at the group level only if the discriminant function combines the predictor X (in this case SES) with information unrelated to the dependent variable. This allows random grouping as a limiting case. So much, then, for the attempt to infer a causal grouping effect from a difference between the two coefficients.

It is a little less obvious that a zero difference in coefficients can arise in the presence of a group-caused effect. For the sake of argument I specify the grouping variable: students are assigned to schools strictly on the basis of SES. If we continue to suppose that group context has no effect, $\beta_b = \beta_w = \beta_t$, since $\phi = 0$. Now suppose that a causal effect of the Coleman type is added. In high-SES schools aspirations of the student body as a whole are given a boost; in low-SES schools aspirations are lowered. This context effect raises β_b but does not imply a change in β_w. In consequence, β_b exceeds the original β_t. This seems to fit with the common reasoning of sociologists. But Meyer among others has suggested another type of context effect: A student's aspiration tends to be raised when he finds himself superior to his schoolmates, and lowered when he is below the mean. This has the effect of raising β_w above the original β_t. Hence effects of the Coleman and Meyer types may offset each other and leave $\beta_b = \beta_w$. Also, a group-caused effect may offset a demographic effect. This reminder of the argument developed at p. 2.6ff. indicates that the absence of a causal grouping effect does not follow from $\beta_b = \beta_w$.

It should be noted that nothing in the model itself implies the presence or absence of a causal effect from grouping. Consider the possibility that groups are formed prior to the events that determine Y ; so η_X^2 is predetermined. Let one collection of groups be treated as groups. Let members of equivalent groups be treated as individuals. If grouping has a causal effect, η_Y^2 and ρ_{XY} will almost certainly not be the same in the two data sets, hence the β_b , β_w values will not be the same. But the formulas of Figure 3.5 fit both studies. This in itself implies that there is no way to reason post hoc about the effects of grouping, on the basis solely of Y-on-X regressions or similar data.

This discussion takes on added importance in the light of Alwin's paper. Alwin shows that the analyses preferred by the Columbia group of sociologists and the analyses preferred by Duncan, Hauser, and others of their persuasion lead to identical conclusions, in the multivariate as well as the univariate case. The fact that $\beta_b - \beta_w$ cannot be directly interpreted implies the need for more elaborate models and for the collection of more evidence on the presumed mediator of the context effect. In the Meyer case, for example, evidence on self-perceived ability would not be hard to collect.

no p. 3.20

Implication for ATI research. I started this investigation with a concern for contrasting regression slopes across treatments. If, I said, the within-group-within-treatment slopes were the same across treatments, and the between-group slopes were different across treatments, this suggested an ATI at the group level (i.e., a causally interpretable group effect). Such a commonsense view must be modified to recognize what has just been said about aggregation effects.

To glimpse some of the problems, let us assume that the relation of Y to X is strictly individual within each treatment (i.e., that grouping has no behavioral consequences). Assume also that the two Y-on-X regressions have an identical positive slope β_t for all individuals pooled. But suppose that the grouping principle used in one treatment differs from that used to form groups in the other. Then, of course, any analysis of between- and within-groups information refers to different populations of groups, even if the individuals came from the same population of individuals. And any difference found in comparing statistics may be attributable to the grouping rule.

NO ¶ Suppose that in one treatment groups are formed by random sampling. Then, within the limits of sampling error, $b_b = b_t = b_w$.

Suppose that, in the second treatment, group formation is influenced by some variable (other than X) that predicts Y well. Then $b_b >> b_t > b_w$, and strong ATI effects will be reported at the between-groups and within-groups levels! It would be possible to concoct an example in which there was no interaction between or within groups, but an overall interaction did appear.

If an ATI study sets out to compare two treatments that are already in place, how groups were formed is crucial to the interpretation. Groups served by one program may differ in their demographic makeup from those served by the alternative program, even though both sets of groups cover about the same range of individual differences.

Perhaps this section can be summed up simply by saying that interpretation of regression coefficients must be exceedingly circumspect when grouping rule is confounded with treatment, so that each treatment is observed in a different population of groups.

Aggregation effects with multiple discriminants. A general formulation can be offered to replace the simple one used to this point. Consider two discriminant functions W and W' . Persons are grouped by imposing assignment rules on the joint distribution of W and W' . The corresponding group means will be denoted by Z and Z' . All persons within a segment are assigned the same Z and Z' . It is easiest to conceive of slicing first on W and then on W' . (If W and W' fell in the X, Y plane, this would divide the original distribution into lozenges.) It is not necessary to make the division by sharp cuts, however. As before, the model applies to data where the investigator did not control assignment to groups; the role of W and W' is to provide a sufficient simulation of the grouping process that might have occurred.

I require that the group means $\overline{X}$ and $\overline{Y}$ be linear functions of Z and Z' , i.e. that $\rho^2_{\overline{X}\cdot ZZ'} = \rho^2_{\overline{Y}\cdot ZZ'} = 1$. This is only slightly restrictive. Given any distribution of $\overline{X},\overline{Y}$ one can easily form a Z and Z' that will reproduce the first and second moments (but not necessarily higher moments) of the distribution. (E.g., let $Z = \overline{X}$ and $Z' = \overline{Y} - \beta_{\overline{Y}\overline{X}}\overline{X}$. Many alternative pairs of continuous W and W' will generate this Z and Z' respectively.)

This model is sufficient to reproduce first and second moments when grouping was regulated by complex contingency rules, since any such rule still leaves us with an $\overline{X},\overline{Y}$ pair for each group, hence a Z,Z' pair. The model would need to be extended to deal with multiple-regression problems.

The development could be stated for any Z and Z', but no information is lost if Z and Z' are rotated within the plane. I therefore work with variables Z_1 and Z_2 which are orthogonal; I require that Z_2 have no correlation with Y·X .

Instead of using partial covariances as in Figure 3.5, Figure 3.7 proceeds more directly to the results. Figure 3.7 starts with a factorial model in four dimensions, with each variable assigned unit s.d. The multiple-regression equations for predicting X and Y from Z_1 and Z_2 are formed. Since all members of a group have the same values of Z_1 and Z_2 , these regression equations predict the $\overline{X}$ and $\overline{Y}$.

Despite the more complex model, the formulas match the results in Figure 3.5, when those are stated in terms of $A (= \rho_{XY})$, tan ϕ and η^2_X . The only functional difference is that $\eta^2_X = C^2B^2 + D^2$, not C^2B^2 as before. Figure 3. 6 applies to β_b but not to β_w , since holding C constant does not simplify adequately. A figure can be developed holding η^2_X and C constant, allowing B to vary. (B implies ϕ).

$$\begin{array}{lllll}
 & I & II & III & IV \\
X & 1 & & & \\
Y & A & +\sqrt{1-A^2} & & \\
Z_1 & CB & C\sqrt{1-B^2} & \sqrt{1-C^2} & \\
Z_2 & D & & E & \sqrt{1-D^2-E^2}
\end{array}$$

$$\sigma(Z_1, Z_2) = CBD + E\sqrt{1-C^2} = 0$$

$$\hat{X} = \overline{X} = CB\ Z_1 + D\ Z_2$$

$$\sigma_{\overline{X}}^2 = \eta_X^2 = C^2B^2 + D^2$$

$$\hat{Y} = \overline{Y} = (ACB + C\sqrt{1-A^2}\sqrt{1-B^2})\ Z_1 + AD\ Z_2$$

$$\sigma(\overline{X}, \overline{Y}) = A\,C^2B^2 + C^2\,B\sqrt{1-A^2}\sqrt{1-B^2} + AD^2$$

$$\beta_b = \frac{\sigma(\overline{X}, \overline{Y})}{\sigma^2(\overline{X})} = A + \frac{C^2B\sqrt{1-A^2}\sqrt{1-B^2}}{C^2B^2 + D^2}$$

$$\begin{aligned}\overline{X} = {} & CB\ (\ CB\ I + C\sqrt{1-B^2}\ II + \sqrt{1-C^2}\ III\ ) \\ & + D\ (\ D\ I + E\ III + \sqrt{1-D^2-E^2}\ IV\ )\end{aligned}$$

$$\overline{X} = (C^2B^2 + D^2)\ I + C^2B\sqrt{1-B^2}\ II + \ldots.$$

If we write X^* for the projection of $\overline{X}$ into the I, II plane, and define ϕ as the angle between X^* and X, $\tan\phi = C^2B\sqrt{1-B^2} / (C^2B^2 + D^2)$

Hence

$$\beta_b = A + \sqrt{1-A^2}\tan\phi$$

$$\beta_w = \frac{A - \sigma(\overline{X}, \overline{Y})}{1 - \sigma^2(\overline{X})} = \frac{A - A\,C^2\,B^2 - C^2B\sqrt{1-A^2}\sqrt{1-B^2} - A\,D^2}{1 - (C^2\,B^2 + D^2)}$$

$$= A - \frac{C^2\,B\sqrt{1-A^2}\sqrt{1-B^2}}{1 - (C^2\,B^2 + D^2)} = A - \frac{\sqrt{1-A^2}\ \tan\phi\ (C^2\,B^2 + D^2)}{1 - (C^2\,B^2 + D^2)}$$

$$= A - \sqrt{1-A^2}\ \frac{\eta_X^2}{1-\eta_X^2}\ \tan\phi$$

Figure 3 .7. Regression coefficients as a function of parameters derived from two grouping variables.

As before, $\beta_b - \beta_w = [\sqrt{1-A^2} \tan \phi](\frac{1}{1-\eta^2})$.

If $\eta^2 = 0$, β_b is indeterminate. If $\eta^2 = 1$, β_w is indeterminate. Disregarding those cases, $\beta_b - \beta_w$ has the same sign as $\tan \phi$.

Assume that ρ_{XY} and ρ_{XZ_1} are positive; this only polarizes those variables. Then $\tan \phi$ can be negative only if $\sqrt{1-B^2}$ is negative, that is to say, variable Z_1 is negatively correlated with $Y \cdot X$. This can arise from a causal effect that places high-X groups at a disadvantage (including a Meyer-type context effect). If

there is no group-caused effect, the negative value can arise from an assembly rule such that groups containing more high-X persons tend to contain persons who are low on some other predictor of Y . For example, if pupils were assigned to classes on the basis of IQ (which can be interpreted as a function of MA - Age), the highest group will be high on MA and low in Age, on the average. Then if MA is used to estimate or forecast achievement, these conditions would make $\beta_b < \beta_w$. Grouping on degree of "underachievement" might produce a similar anomaly. It appears unlikely that demographic effects alone will often make $\beta_b < \beta_w$.

The relation $\beta_b = \beta_w$ occurs with grouping on some combination of X with an irrelevant variable (perhaps a random assignment process).

This requires that there be no demographic variable or other precursor X' such that $\phi \neq 0$.[4] That is, X completely specifies the relevant grouping variables.

Notes for Section 3

p. 3.1 [1] Walberg (personal communication) suggests that the mean should not be used as an aggregate statistic because of its sensitivity to skew and especially to outliers. Decker Walker suggests that the model should provide for nonlinear regression from the outsets (and this does appear to be important with a categorical variable such as that of Bowers). I prefer to leave these possible elaborations in the background. Investigators should inspect plots of between-groups and pooled-within-groups relations.

At different places in this report I have shifted notation, more because the sections were drafted at different times than for any good reason. What appears here as μ_{X_c} was simply $\overline{X}$ in Section 1 and X_c in Section 2.

p. 3.2 [1a] Others have used E^2, the correlation ratio, in place of η_X^2. When class membership is fixed, as I assume, the two are identical.

p. 3.14 [2] When Z coincides with X and each differential element of Z defines a new group ($B = \pm 1$, $C = 1$, $\phi = 0$), the variance of X within groups is zero and the slope is undefined. With $B = \pm 1$ and C even slightly different from zero, the within-groups slope becomes A.

When Z coincides with $Y \cdot X$ or projects into $Y \cdot X$ ($\phi = 90°$), the between-groups variance is zero and the between-groups slope is undefined. As ϕ increases toward 90° the slope becomes indefinitely great; as ϕ decreases from 0° toward -90°, the slope becomes indefinitely great but negative.

Statements about η_X and η_Y in Figure 3.5 are true when both are given positive signs.

p. 3.17 [3] Personal communication indicates that Duncan does not wish to defend the argument discussed here; it was formulated nearly ten years ago. Today he would emphasize that coefficients are highly equivocal unless there is commitment to a causal model of the process of group formation and of the generation of the dependent variable. The Alwin paper indicates that the reasoning of the 1972 publication has not been superseded in the sociological literature.

p. 3.24 [4] With a finite number of groups a strictly random process may generate a value of ϕ that is far from zero in either direction.

4 . The reference population and its parameters

Alternative models for statistical inference

Data on students observed in a group of classes could be interpreted with no attempt to generalize. That is, the classes, and the students within the classes, could be regarded as fixed. (One could consider the data themselves as a sample of observations that might have been made on these same subjects, in which case one would generalize over the universe of observations.)

The most obvious way to frame a generalization is to assert that persons and classes are randomly sampled from a population of students. This requires drawing students randomly and independently to fill each class in turn, which would make approximately zero the intraclass correlation for every initial characteristic of the persons. This is not reasonable for most groups that exist in society, and it is likely to be contradicted by the data in hand.

In trying to identify more plausible alternatives, I confine attention to two levels, collectives and members. The ideas apply to subcollectives as much as to individual members, and are readily extended to additional levels. I assume that it is intended to generalize over collectives, and that the collectives are a random sample of the population of collectives. Collectives, then, may be considered "random" as that term is used in the statistical literature.

In deciding whether to treat members as fixed or in some sense random, the key lies in the structure of the population. Are all the collectives separate and distinct? Or do different sets of members constitute realizations of "the same collective"? The second alternative applies most obviously when the population of collectives extends over localities and over time. High-school student bodies have different members each year. An investigator interested in persistent differences among schools might think of the population as comprising a number of "local" populations, each made up of a succession of student bodies in different years. (The population may be finite.) Classes within a population of classes might likewise be identified with teachers, so that the potential members of classes of one teacher constitute a local population. I next discuss the model for inference that follows from each of the alternatives. At the end, it will be possible to discuss the bases for choice between the two conceptions.

Collectives distinct, persons fixed. In the first formulation, collectives are regarded as without connecting identities, just as persons are in the usual models for inference. Collectives are sampled independently, from a population of collectives that might have been formed by applying a particular grouping rule to a population of individuals. The grouping rule may be under deliberate control or may be a social process that can be only inferred from the data.

It seems to me that under these circumstances members have to be regarded as fixed. A certain group of persons was assembled and together went through certain events. Those events constitute a unique history. There is no basis for speculating as to what would have happened if, at the outset, Billy had been replaced on the class rolls by Milly. What went on in the class may have been influenced by the synergism between Billy and certain others; to have enrolled Milly instead would have made the class a different experimental "subject". The class containing Milly might have been drawn under the grouping rule, but it is a distinct class and only one of numerous alternatives to the class observed. The model does not allow for a close family tie of the class with Billy to particular other classes in the population, except as classes may be blocked a posteriori on selected aggregate and global variables.

With students fixed, effects within the classroom have to be looked upon as historical accounts of the consequences of bringing together this set of students, this teacher, and whatever unpredictable events affected the group. The unforeseeable variability in delivery of the treatment, in classroom morale, in epidemic illness, etc. is a part of the causal history. As a thought experiment one can ask what would have happened if this same collective had gone through the specified treatment several times independently. That is, one can be conscious that fortuitous events played a role in determining the history and the scores. But there is no satisfactory way to assess such variability. The one empirical approach is to treat successive units of instruction in the class as independent events, but even if the topics are unrelated, the first experience is likely to influence the second. It _is_ practicable to generalize over the universe of observations of the outcome -- but that is a side issue here.

Collectives nested within local populations. The alternative recognizes the division of the population of members into what can conveniently be called local populations. The grouping rule determines the membership of each local population. Random definition of local populations is unlikely. Each local population is a subpopulation of the population of collectives defined by the grouping rule. It consists of all the collectives belonging to a certain locality -- i.e., all the "remedial" sophomore English classes that might be formed in this school during a 10-year period. Here collectives are simply nested within localities. One might think of crossing localities with time periods and identifying a place-time combination with a subpopulation.

Student bodies, neighborhoods, and classes are not formed randomly, as is evidenced by the usual intraclass correlations on initial characteristics. But it seems not too unrealistic to assume that any one collective within a local population is a random sample from a set of collectives that might have taken its place. Thus, if local populations of classes are identified with teachers, where each teacher has many potential classes, I allow the intrateacher correlation on an initial variable to depart from zero, and assume that the intraclass correlation within a teacher fluctuates around chance expectancy.

Even with this model, sometimes it is appropriate to regard the collective observed as having a fixed membership and generating a unique history. Then one could evaluate the relevance of collective-level statistics to the subpopulation only by observing two or more collectives from that subpopulation.

The independence assumption.

When the investigator chooses instead to regard the members as random, he is making two assumptions. He is assuming that one member of the local population has as good a chance to fall into the sampled collective as another, which seems plausible. Second, he is assuming that as the events of the treatment period unfolded, each member's history and performance developed independently of the experiences and acts of his classmates. This seems more likely to fit the facts of individual instruction than of group instruction. But let me be more precise.

I elaborate on the model of Section 3. Assume a population of subpopulations for which there is a single β_b and β_w. There is corresponding population of values of μ_{X_c}, γ_c, and $\beta_c - \beta_w$, one value of each for each subpopulation. Here, μ_{X_c} is the expected value of X_{p_c} over members of subpopulation c. Second, there is a grand population of deviation scores for members, $X - \mu_{X_c}$ and $Y - \mu_{Y_c}$. ($\mu_{Y_c} = \beta_b\mu_{X_c} + \gamma_c$.) The variances of the means are a function of the intraclass correlations.

With this model of the population, the logical design is to sample local populations and then, to represent those chosen, to sample one or more collectives. The calculated $\hat{\gamma}_c$ and $\hat{\beta}_c - \hat{\beta}_w$ for a particular sampled collective estimate corresponding parameters for the subpopulation.

¶ To evaluate the sampling error when only one collective per subpopulation has been observed, one uses the member as unit of sampling and estimates the variability of the mean or regression coefficient from the within-collective variance.

This amounts to viewing the members as independent instruments for observing an effect, an effect that is associated with the subpopulation no matter which members constitute the collective. The obvious example is a teacher effect. The teacher may be supposed to generate an effect of size γ_c in every class, by virtue of excellent (or poor) technique. Individuals affect the mean on Y through their aptitudes, but if the model is properly specified that contribution is separated from γ_c. Likewise, the teacher may adopt some tactic, such as fostering competition, that generates the same $\beta_c - \beta_w$ in every class he teaches. The independence assumption fits well with some conceptions of teacher effects, school effects, and context effects generally.

It is not hard to generate counterexamples, starting with contagion effects. Most teachers have the impression that classes develop their own "personalities" -- responsive, recalcitrant, mutually supportive, divergent, etc. This implies a variability across classes within the subpopulation much greater than one would estimate from the within-class variation. On the other hand, consider the teacher who "grades on a curve", so that every class has almost the same final rating. Then the variation of mean ratings across classes will be much less than one would estimate from variation within the class.

Choice among models. The first summary comment to make is that, for many research purposes, inference regarding members of collectives is of secondary importance at most. In compensatory education, the chief question is whether, on the average over (presumably) districts, one policy is more profitable than a competing policy. An experiment on instructional method usually seeks a conclusion that can be generalized over classes or possibly teachers. For such purposes, the collective is the unit of decision making or of theory, and it seemingly should be the primary unit of sampling, assignment, and analysis. In such a context, however, an investigator might appropriately make supplementary studies of classes, asking why some have large means or large regression coefficients. At this point, he does face a choice between regarding the class statistics as representing a fixed history or as representing the independent histories of its members.

To think of members as random and independent appears to require, first, an identification of local populations. To simply say "pupils are regarded as random" (for example) is to make a deniable assumption of zero intraclass correlation. In the kinds of studies this report discusses, local populations are readily conceived, so that is no barrier. The point is primarily important in stressing that the variance over members in pooled collectives is not a proper basis for inference; the model directs attention to variance of members within collectives. Inference based on this variance has to do with parameters within local populations, not with inference over the undifferentiated grand population.

Second, a substantive decision is required as to the legitimacy of the independence assumption. Bowers might want to set confidence limits on the mean proportion cheating in College A, over its subpopulation of successive student bodies. He surely would not regard students as independent; his very hypothesis regarding the context effect seems to imply a positive feedback loop among the members. Some years, then, are likely to see more cheating than others, and within-year variance of students would tend to give a conservative confidence interval. An investigator evaluating programmed instructional materials might be convinced that students react to the materials independently, so that each one earns about the same score as he would have if taught alongside other classmates. Then he can contentedly regard students as random while estimating the extent to which certain teachers get superior results. (Students are random within the subpopulation for the teacher, however.) Third, think of research like Barker's, where the local population is the community and the variable of interest is the number of responsible tasks undertaken in the community. The variability over persons within the cohort will give too wide a confidence interval for the community mean. The number of roles to be filled is finite, hence the mean over cohorts must be quite stable; yet there is a large variance within the cohort.

An investigator who has only one sample from a subpopulation and wants to infer to the subpopulation must develop a substantive argument about the direction and amount of bias the independence assumption entails. He may go on to base an inference on this shaky assumption, with appropriate caution.

In this report I have chosen to regard members as fixed within collectives. This is a conservative position that limits the number of issues I have to deal with in each particular example.

Weights that define parameters

Population parameters have to be defined with due regard to the number of members per collective, when this is not constant. Parameters may weight by the number of members or may weight collectives equally. This requires a conscious choice of the parameters to be estimated. To be sure, a person who is interested in the weighted mean for school districts may use the unweighted mean as an estimator, assuming that the correlation of the variable with district size is negligible. But the fundamental question is what mean (or variance, regression coefficient, etc.) he would like to evaluate. Sometimes the decision can be reduced to a theoretical question and sometimes it is a question of utility. The same weighting should, I think, be used in defining all the parameters of the study, to avoid numerical inconsistencies.

I am inclined to think that in most instructional research weighting pupils equally is the preferred way to define parameters. It is doctrine in our society that individuals are equally important, and in any ultimate policy decision the burden of proof is on whoever proposes to weigh pupil interests unequally. That is, if it should happen that the mean effect of a treatment is positive when calculations are weighted by class size, and negative when unweighted, "the good of the greatest number" would favor use of the treatment. Weighting classes by n_c weights individuals equally.

Theory may give a reason for weighting on a principle other than "one man, one vote". In research on factors influencing national returns for Senate seats, the fact that each State has two Senatorial votes might argue for using unweighted State means. This, it should be noted, arises not from a statistical principle but from a substantive context in which States are equivalent in weight. Weighting groups equally can be appropriate in education also. One example is an evaluation study with wear-and-tear-on-the-teacher as dependent variable. Teachers, in that instance, are the ones with equal rights.

When the State of California wants to examine the mean of student achievement, it might count districts, or schools, or pupils equally. It seems obvious that pupils are the correct unit. If a change in distribution of tax revenues depresses the school program in 20 large cities while improving the program in the 500 smallest districts, the effect on the welfare of pupils qua pupils probably is negative on balance. Surely the legislature is just as concerned with the welfare of the typical city child as with that of the typical child in a small community.

With regard to the regression of district mean outcome on background characteristics, if district size makes much difference there probably should be a separate regression in each size category. But if only one regression is to be used, the pupil-weighted regression for district means seems to give the best statement as to the "normal" district-level outcome corresponding to given background characteristics.

Consistently weighted calculations produce harmonious numbers at several levels. For example, the pupil-weighted sum of squares between districts and the sum of pupil-weighted within-district sums of squares for schools add up to the pupil-weighted sum of squares for schools pooled. Inverse weighting is equally possible. If one wanted to weight districts equally in district-level calculations, it would be possible in school-level calculations to weight each school in inverse proportion to the number of schools in its district.

The weighting that defines a parameter may not be the weighting used in making estimates, particularly if the sampling fraction varies with the collective. One might, in California, sample schools in large districts while collecting data in every school in the small district. This would lead one to weight schools unequally in calculations over districts.

In Bowers' study, data on attitudes and conduct were collected in 93 colleges. The same number of students were sent questionnaires in each college, though returns were not uniform. The sample sizes were not at all proportional to the college enrollments. The population of interest could be defined by

a. counting respondents equally, or

b. counting colleges equally (which would call for weighting each sample mean equally and in individual-level calculations weighting the data inversely by the size of the sample for the person's college), or

c. counting individuals in the national student population equally (which would call for weighting each datum by the ratio of college enrollment to sample size for that college).

Bowers, it will be recalled, was concerned with the relation of behavior to the dominant opinion in the student body. This is described in the between-colleges regression and in the mean of the within-college regressions. Option (a) -- which Bowers and others used in their calculations -- seems not to be the soundest choice. The resulting statistics refer to no population save that constituted by the sampling procedure. Options (b) and (c) could give disparate results if means for large and small colleges are not distributed around the same regression line, or if their within-college regressions differ systematically. If the large colleges exhibit a positive trend and, *arguando*, the small ones exhibit a negative trend, these can balance out in the unweighted calculation whereas the large-group trend will dominate the weighted calculation. I believe that Bowers would be interested in a trend whether it appears in the weighted or the unweighted calculation. In investigations like this, as in the California data, it appears important to learn how regressions vary with group size.[1]

If Bowers were to decide that size was not systematically related to the effects of interest, he might want to take the precision of his information into account in estimating the relationships. If student bodies are much larger than his samples, the standard error of each college mean is nearly inversely proportional to the square root of the sample size for it, and the means could be weighted by that factor in the between-groups analysis. The same weighting could be used in averaging the within-college regression coefficients.

Illustrative statistics for populations of collectives

The population of collectives, I have said, is characterized by a number of parameters at the level of the collective. Two examples will give concreteness to the idea.

Head Start. Smith and Bissell (1970) give correlations, means, and s.d.'s for a set of demographic variables and a posttest (Metropolitan Reading Readiness) on 202 Head Start children in 26 centers. The entries in Table 4 .1 are calculated from their report. As the data come from a sample of centers they describe the reference population only approximately. The covariances of the initial variables are as much a part of the population definition as are the variances. In fact, Smith and Bissell described the data in such detail in order to point out that the Head Start group matched the control group poorly. Even though the means and standard deviations matched fairly well, the correlations were consistently stronger in the control group. POPED and NKIDS, for example, have a covariance of -0.39 in the Head Start sample, and -0.77 in the control sample (between centers).

Table 4.1. Statistics describing a sample of Head Start centers

		Between-centers variances and covariances						b with Reading (between centers)
Variable	Mean	POPED	POPINC	POPOCC	NKIDS	MR	η^2	
Father's education (POPED)	2.3	.36					.31	3.47
Father's income (POPINC)	2.6	.16	.49				.43	6.06
Father's occupation (POPOCC)	1.0	.07	.20	.25			.20	6.72
Children in family (NKIDS)	4.8	- .39	- .02	.02	1.0		.20	.72
Metropolitan Reading (MR)	52.2	1.25	2.96	1.68	- .72	75.69	.29	

School districts in California. Another example comes from the California Assessment Program. Every student in certain grades is tested each year. Rogosa and I have analyzed data for 882 districts (4514 schools); this is not the entire population, since we confined the analysis to schools for which information was available on each of the variables under consideration. These were:

ELT 3.	A readiness test given to first-graders entering in 1973.
ELT 4.	A similar test given to first-graders entering in 1974.
Rdg.	A reading test at end of third grade, given in 1975 to students most of whom entered in 1972.
SES.	An estimate based on teacher's report of father's occupation for each third-grader.
Mob.	Principal's estimate of per cent mobility for the school.
Bil.	Teacher's estimate that the pupil was or was not bilingual.

Calculations can be made directly from school means and from district means, but in the district-level calculations we weighted by number of schools. (In retrospect, we had better grounds for weighting by number of pupils.)

Table 4 .2 gives results for all districts except those having just one school. The correlations are large for all variables except mobility. The intraclass correlations are larger than in the Head Start data and, except for mobility, remarkably uniform.

Even in this weighted calculation the elimination of the one-school districts had a large effect. The ELT-3 vs. ELT-4 correlation dropped from 0.95 to 0.84. No interclass correlation increased. The standard deviations for schools did not change but those for districts increased. Consequently, the intraclass correlations rose to about 0.50 (0.40 for mobility). To at 4 .12

Table 4.2. Population parameters for California districts with more than one school. All calculations weighted by number of schools per district.

				Between-districts correlations					
Variable	Mean	s.d., schools pooled	s.d. for districts	ELT 3	ELT 4	SES	Mob	Bil	η^2
ELT 3	29.04	2.18	1.41						.42
ELT 4	27.45	2.49	1.64	.95					.43
SES	2.16	0.41	0.27	.79	.80				.43
Mob	39.56	11.97	6.18	-.16	-.16	-.28			.27
Bil	0.18	0.19	0.13	-.75	-.74	-.61	-.02		.47
Rdg	82.34	9.18	6.02	.89	.89	.81	-.23	-.67	.43

analyze the one-school district separate from the remainder makes sense; and in the population of larger districts it is advisable to check that relations of interest do not vary with district size.

Los Angeles (441 schools) is four times as large as the next largest district, which led us to wonder whether Los Angeles alone had an appreciable influence. The correlations with Los Angeles omitted departed little from those in Table 4 .2. The s.d.'s for schools decreased by about 10 per cent (except for SES and mobility) and η^2 increased. Removing Los Angeles had little effect on most of the statistics because its mean was close to the State mean. Had it been an outlier on any variable the changes would have been great.

A problem of estimation

If one wishes only to describe relations in the sample of groups and individuals before him, it is unnecessary to speak of "estimation". Calculation simply requires attention to the definition of the various components and parameters, with respect to such matters as weighting. I postpone most problems of inference to Section 7. One point needs to be made here, however, to prepare the reader for the erratic behavior of the between-groups coefficients to be encountered in Section 5.

A regression coefficient is determined largely by the cases toward the extremes on the predictor variable. Those cases "have leverage" on the slope of the regression line, just as do persons perched on the end of the seesaw. Cases near the mean -- the fulcrum -- have little influence on the slope. This means that the "effective" sample size determining a regression coefficient is much less than the number of sampling units.

4.11 Note 1. We made some limited comparisons in some of the Bowers data and found that regressions were similar whether colleges or individuals in the sample were weighted equally. We did not apply weighting of type (c).

5. Illustrative ATI studies

The Anderson study

Before considering theory further, I turn to a number of illustrative studies, beginning with G. L. Anderson's 1939 data. Webb and I reanalyzed that study because the ATI effect it reported has been of considerable interest. A full account of the design of the study and of our reanalysis appears in the Cronbach-Webb (1975) paper, so I can be brief here. Data on 9 classes in Treatment A and 8 in Treatment B are available. The classes were taught the same year-long arithmetic curriculum -- the A's by a method that emphasized the meanings of the processes, and the B's by a drill method, with little meaning being developed. Teachers were assigned to the method most like their usual style, not randomly. The students in each class were those assigned to that teacher by the school's routine procedure. The study is a quasi-experiment. One can reasonably generalize from the A data to the population of teachers likely to opt for a meaningful method (in the schools of the late 1930's). I prefer not to regard the A and B teachers as samples from the same population. The classes may well be random samples from a single population of classes, but classes within treatments differed in ability level.

Among Anderson's many pretest scores, we found it sufficient to use just two, which we label ABILITY and PRECOM. The former is a conventional group mental test rescaled to have mean zero and s.d. 100 over all cases pooled. The latter is the total score on the Compass achievement test in arithmetic computation, at the time of pretest. It too was put on the 0,100 standard-score scale. The dependent variable (ZACH) was a similarly scaled composite of subtests from the Compass posttest and from the Analytic Skills of Attainment. Rescaling makes it easier to compare regression coefficients for variables with different metrics. These will not be standardized regression coefficients. The s. d. of each variable varies with the group.

Webb and I did not use the single stepwise regression analysis suggested at p. 3.2, because it is a comparatively awkward way to arrive at descriptive statistics for separate classes. Instead, we carried out separate **regression analyses** within treatments and within each class. This costs more in computer time than a single generalized analysis, but the ease of interpretation saves investigator time. The procedure does not, however, generate inferential statistics on the treatment contrast.

A weighting decision. To evaluate $\hat{\beta}_b$ and other between-group statistics within a treatment one has these options:

1. Calculate μ_{X_c} and μ_{Y_c} for each group. Enter these k pairs in the computations. Or
2. Carry out the computations but weight each pair $\{\mu_{X_c}, \mu_{Y_c}\}$ by the corresponding n_c .

In the model Webb and I used η_c as a weight in defining parameters. Weighted calculations from the sample give unbiased estimates of the weighted parameters and the unweighted calculations in general do not.

Regressions of ZACH on PRECOM. Within each treatment, regression analyses were made with the group mean on PRECOM as predictor and with the individual's deviation score as a predictor. I ignore the constant terms, which are of no immediate interest in this report; the treatment means did not vary greatly. The unstandardized regression coefficients were as follows:

	Between classes	Within classes
Drill treatment	0.74	0.73
Meaning treatment	0.47	0.71

¶ The difference in the between-classes coefficient is large enough to be of potential theoretical and practical interest *if* taken at face value. Apparently, differences in X means produce comparatively large differences in outcome means of drill classes. One can rationalize this by hypothesizing that when an able class shows good results on drills the teacher steps up the pace and covers more topics or more variants within a topic. Increasing (or reducing) the amount of work covered is comparatively easy in a drill class. Practically, this difference in coefficients coupled with a near-zero difference in overall means suggests the hypothesis that the drill method is best for classes formed of high-PRECOM students, and the meaning method best for low-PRECOM classes; but this would require verification on classes formed in that way.

A between-classes coefficient in a small study cannot be trusted. Even though Anderson's study was large by conventional standards -- over 400 students -- only 8 or 9 classes contributed to each regression equation. A difference in coefficients much greater than 0.27 (in this metric) would fall short of significance with such a sample. When we plotted the means (see figure) the two sets of points seemed to lie within the same distribution. Coefficients are most strongly influenced by data points at the extremes of the X scale. At the right end the extreme points in the two treatments are close together. At the left end, the extreme point for drill pulls its slope down, whereas the extreme point for meaning is very little below zero on the Y scale. This alone seems to produce the difference in final slopes. For a more formal consideration of statistical inference, see Section 7.

The within-class coefficients are almost exactly equal. Taking the coefficients at face value, the two coefficients for drill are the same, which is consistent with the view that individual aptitude determines performance and context effects are lacking. These data, however, give no basis for ruling out the hypothesis that if students were taught individually the overall slope would become much flatter (no systematic adjustment of the pace to ability) or much steeper (students truly moving at their own rate). The fact that the between-class slope is smaller than the within-classes slope for meaning would invite other speculative interpretations -- for example, that the comparatively able members of a class drive the level of discussion up to the point where the less able become confused. All such interpretations become moot when the uncertainty attached to the between-groups coefficients and the possibility of demographic effects are borne in mind

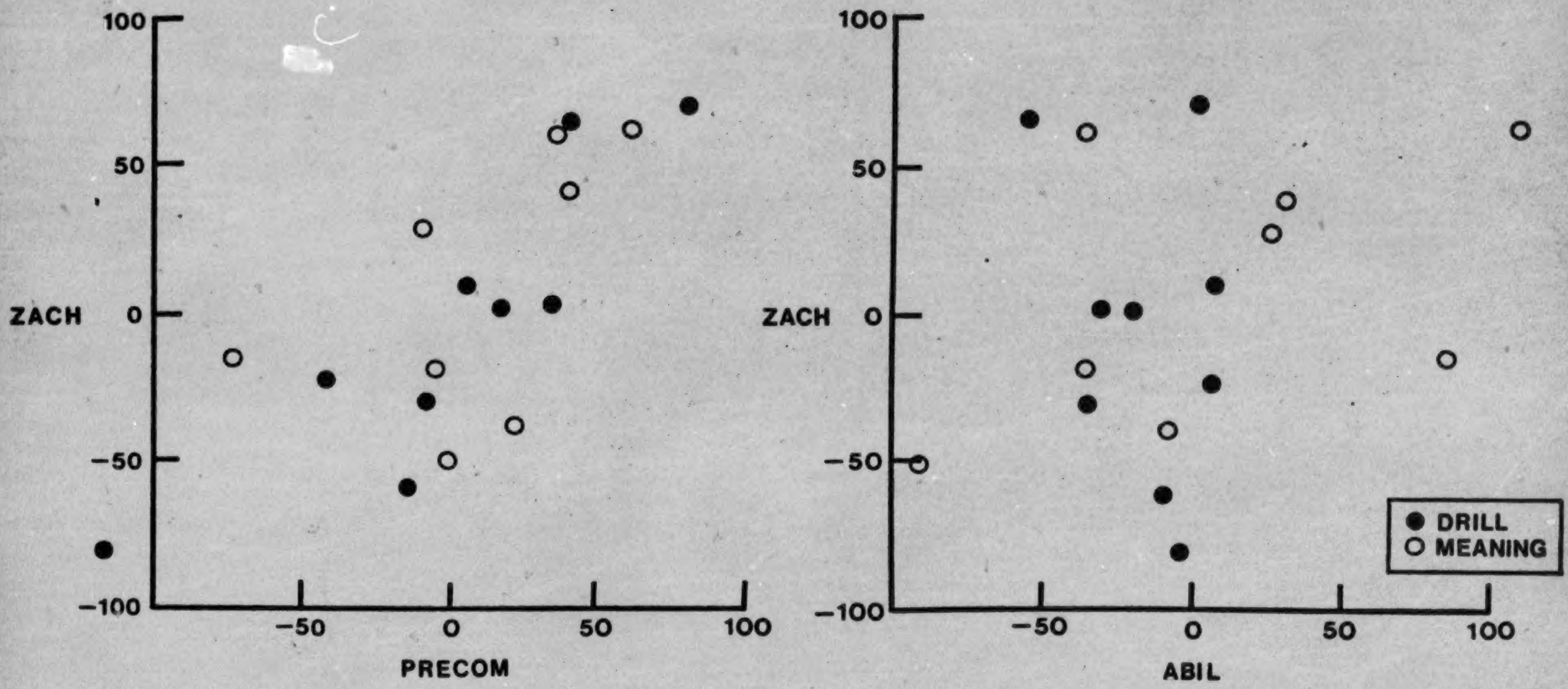


Figure 5.1. Plot of class means in the Anderson study

(From Cronbach & Webb, 1975)

at 5.4

Along with the coefficients within single classes, the following array gives the class means on PRECOM in parentheses.

Drill:	1.07	.93	.90	.90	.84	.63	.56	.55	.51
	(34)	(-45)	(38)	(17)	(-7)	(6)	(-13)	(81)	(-118)
Meaning:	1.12	1.08	.87	.76	.73	.67	.50	.43	
	(-1)	(-7)	(38)	(23)	(-75)	(-8)	(108)	(36)	

These distributions of b's do not differ. The weighted averages are 0.75 and 0.78 respectively.

An investigator gathering data such as these today would be wise to ask why some coefficients are twice as large as others. A coefficient is an historical fact about a certain group of identified students, going through a unique series of local events that realized in a specific way an intended treatment plan. It is as legitimate to contrast high- and low-slope classes as it is for the historian to contrast, say, utopian settlements that succeeded with utopias that failed.

In the drill classes, the β_c's are positively related to the class means ($r = 0.28$). The low-slope outlier whose mean is -118 contributes so much to the correlation that it probably should not be taken seriously. In the sample of meaning classes, the correlation is -0.34; the outlier whose mean is -75 weakens an otherwise-strong negative trend. When the sample of classes is of the normal size, trends such as these will never be convincingly established as characteristic of the population of classes. Nonetheless, the analysis is a reasonable step in learning from the data.

The variance of ZACH scores within treatments was broken down as follows (all figures are percentages):

	Drill		Meaning	
Between classes	21.5		14.9	
Regression β_b		16.5		8.0
Residual		5.0		6.9
Within classes	78.5		85.1	
Regression β_w		51.5		57.3
Regression $\beta_c - \beta_w$		4.1		5.6
Residual		22.9		22.2
Individuals (overall)	100.0		100.0	

These values are pretty much what one would expect: within-class differences account for more variance than between-class differences, and the predictable variance is larger than the unpredictable variance. The specific-within-class regressions do not account for much variance. Comparatively little of the between-class variance in outcome under the meaning treatment was predicted; this is consistent with the slope reported earlier.

It is well to keep absolute magnitudes of effects in mind. (In the Anderson data, the ZACH variances within Drill and within Meaning were 9880 and 10011 respectively.) A between-classes effect, for example, should not be dismissed as unimportant merely because it is small relative to the scale of individual differences; at some point, one must consider the meaning in absolute terms. When a test is the dependent variable, what is "important" is judged on the basis of the absolute proficiency required to earn various scores.

Regressions of ZACH on ABIL. Anderson's original analyses were bivariate, and at the individual level. His regression planes relating achievement to ABILITY and PRECOM had different slopes. Drill appeared to generate better achievement for students with high PRECOM and comparatively low ABILITY whereas meaning gave better results for those with the reverse pattern ("underachievers"). Before making this calculation, Anderson removed a subset of superior classes from the sample. Webb and I retained all classes in our calculations.

Instead of analyzing ABILITY we formed a variate ABIL defined as the value of ABILITY - 0.47 PRECOM, restandardized to a 0,100 scale. Since 0.47 was the overall regression coefficient relating ABILITY to PRECOM, ABIL and PRECOM have little redundancy at the individual level. To have used ABILITY as a predictor in a univariate analysis would echo so much of the information in PRECOM as to obscure the interpretation.

The unstandardized regression coefficients onto ABIL were as follows:

	Between classes	Within classes
Drill treatment	-0.20	0.39
Meaning treatment	0.31	0.52

When PRECOM was the predictor, the Anderson finding had led us to anticipate larger coefficients in the drill treatment. This was true only of the between-classes coefficient, and we have dismissed that finding as untrustworthy. Anderson led us to anticipate smaller coefficients in the drill treatment with ABIL as predictor, and again the principal difference appeared between classes. The difference is impressively large -- but is it worthy of serious consideration?

The plot of group means again suggests that the two sets of means have the same distribution, in that range of ABIL where both treatments appear. The negative slope in drill would turn slightly positive if one class at the upper right were discarded. The salient feature of the plot, however, is the narrow range of ABIL means in drill classes.

Tracing this back, we found that across drill classes ABILITY and PRECOM were highly correlated (0.74), but the correlation was near zero (0.09) across the meaning classes. The drill-class means in ABILITY were largely redundant with PRECOM. Consequently, there was little variance in the second dimension of the between-class predictor distribution. The small variance in ABIL across drill classes meant that the between-class slope onto ABIL is almost worthless as an estimate of the population regression.

Anderson's study foundered on an accident of sampling. The classes in the two treatments appeared comparable to him, since the univariate between-class distributions on ABILITY and PRECOM were similar. Also, the bivariate distributions for individuals pooled looked much the same. Anderson failed to inspect the bivariate distributions of class means. The points in the drill distribution lie nearly in a straight line, whereas the meaning distribution is elliptical. A chance failure to assign "off-line" classes to the drill treatment spoiled Anderson's chance to get information on the bivariate regression. Smith and Bissell, it will be recalled (p. 3.6), found a similar anomaly in the between-groups covariances of predictors in the Westinghouse study, even though Head Start and control cases had supposedly been matched. The Westinghouse investigators evidently inspected univariate between-center statistics and, like Anderson, failed to observe the mismatch of the multivariate distribution of center means for predictors.

In Anderson's data, the slope difference onto ABIL within groups is too small to be worth interpreting. I shall not pursue further details of the study.

Cooperative Reading data*

Plan of the studies. The Cooperative Reading study of the mid-1960s was a forerunner of other "planned variation" studies. To compare a dozen methods for teaching primary reading, 27 research contracts were let. Each investigator was to adopt certain features of a standard design, but he was free to add procedures and to introduce treatments that interested him alongside the standard treatments. We concentrate on the comparison of Basal (B) and Language Experience (LE) methods. Each investigator prepared his own reports, and a composite analysis of all the data was made by Bond and Dykstra (1967). The reports attracted our attention because many ATI were reported; a summary of those, prepared mainly by Snow, appears in Cronbach and Snow, 1976, Chap. 8.

The director of each of the 27 studies selected intact classrooms whose teachers agreed to participate in the study and assigned classes to treatments. Directors did some matching of teachers across treatments on the basis of amount of experience, and on achievement of their students in the previous year. In most of the projects teachers ranged widely in rated competence. Most teachers were experienced in teaching first-grade reading using basal readers; few had taught by LE.

Some project directors matched classes across treatments on the basis of student aggregate performance in kindergarten, and on aggregate SES. Most projects used students of varying ability. In some projects, ethnic backgrounds of classes happened to differ from treatment to treatment. In few of the projects comparing B and LE were classes randomly assigned to treatments.

*Noreen Webb is coauthor of this section.

In the Bond-Dykstra analyses comparing Basal to other methods, the non-Basal approaches seemed in general to be superior to Basal programs.

N.9 [Students superior in certain abilities seemed to achieve better in LE than in B. Less able pupils profited more from B. This relation was not clearly interpreted, however, since Bond and Dykstra were unable to carry out a multivariate ATI study using all readiness and aptitude scores together.

A reanalysis of a subset of the data with sophisticated multivariate techniques was made by Lo (1973). He reported a significant advantage for students with high perceptual speed (i.e., high on Identical Forms) in LE, whereas those low on the scale did better in B. Lo's analysis pooled classes and projects within treatments.

The original analysis across projects. Four projects compared B with LE. The B classes usually followed traditional Ginn or Scott-Foresman readers. In LE classes pupils told stories; these stories formed reading material which incorporated the children's language patterns. The methods varied slightly across projects.

Treatment groups within projects ranged from 219 pupils to 652 pupils; the number of classes per treatment ranged from 10 to 27. Class size varied from 8 to 32.

Students in all projects were tested in September of Grade 1 on the Pintner-Cunningham Intelligence Test and on several more specific variables (e.g., Phonemes, Pattern Copying, Word Meaning). Five subtests of the Stanford Achievement Test Primary Battery I were administered after 140 days of instruction.

Bond and Dykstra first analyzed in a Sex x Treatment x Project design, working from the means for boys and girls in each class. The unit of analysis was thus the half-class mean.

Analysis of variance was performed on each pretest or posttest. Two analyses of covariance were performed on each posttest, one with Phonemes and Identical Forms as covariates, the other with all seven pretest measures as covariates.

Girls scored significantly higher than boys on all pretests. On 6 out of 7 pretests projects differed significantly. Treatment groups differed significantly on 4 pretests. On one pretest, a significant

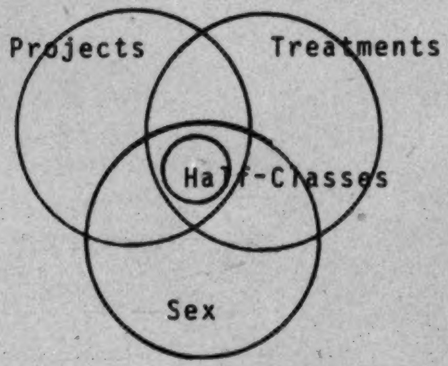
Project x Treatment interaction appeared. These last results strongly imply that the treatment groups are not random samples from the same population, though the significance tests are questionable (see below).

Most sex differences at posttest tended to disappear in the covariance analysis. The few treatment differences found could be attributed to chance.

Significant differences among projects and Project x Treatment interaction effects turned up. Because the treatments behaved differently from one project to another, Bond and Dykstra decided to analyze within each project.

Half-class as unit of analysis? The Bond-Dykstra anova is exceptional in its design, and a discussion of it will extend thinking on units of analysis. The example is so exotic, however, that I give it little space. Their design and analysis are shown schematically in Figure 5.2. Three factors are crossed. Male and female halves of a class were taken as the unit of analysis, with no attention paid to the nesting of halves within classes.

To ignore classes is in effect to assume *a priori* that the component of variance associated with classes is zero. This would be self-evident only if half-classes had been formed and treated separately, and then arbitrarily paired for the analysis.


Projects
Treatments
Half-Classes
Sex

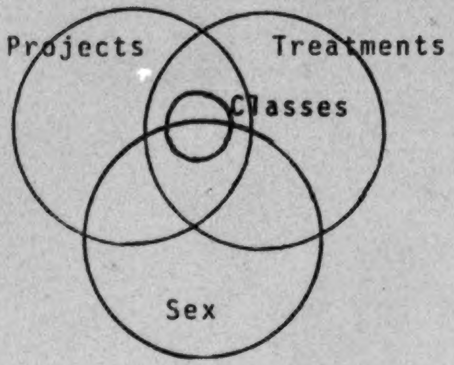
Effects

Projects i
Treatments j
Sex k
ij
jk
ijk
Half-classes within ijk Error term

Figure 5.2. Scheme of the Bond-Dykstra anova

Effects

Effects


Projects
Treatments
Classes
Sex

Projects i
Treatments j
ij
(a) Classes within ij Error term
Sex k
ik
jk
ijk
(b) Half-classes within ijk Error term

Figure 5.3. Analysis

An analysis that more adequately recognizes the pairing of half-classes looks upon the half-classes as repeated measures on the class. (This formulation was suggested to me by Dan Davis.) The analysis proceeds as suggested in Figure 5.3. Since sex is a fixed factor, the error term for the i, j, and ij effects is (a), the mean square for classes. The error term (b) applies to the remaining effects.

The error mean square of Bond and Dykstra is closely related to (b), but it includes variance attributable to class and it claims twice as many d.f. for error.

The original analysis within projects. In the within-project analysis Bond and Dykstra made the student the unit of analysis in order to look at ATI effects. The treatment main effect usually favored LE, though this was reversed in some projects. To study interactions, subjects were blocked in turn on Pintner (4 levels), Phonemes (3 levels), and Letter Names (4 levels). In only one project (Stauffer) were many interactions reported as significant; the child with poorer readiness tended to achieve more in B, whereas the able child profited more from LE. Bond and Dykstra dismissed this result, judging that the apparent interaction effect with one variable might be accounted for by initial differences on other variables. They lacked a procedure for handling the pretests simultaneously.

Bond and Dykstra were too hasty in dismissing ATI in the other projects simply on the basis of nonsignificance. According to Cronbach and Snow, blocking on aptitude produces extremely weak significance tests even when students are properly the unit of analysis and N is large.

Bond and Dykstra attributed to chance several borderline significant interactions that involved the same outcome variable. But when significance tests lack power, it is a mistake to let descriptions of nonsignificant but interesting effects drop from sight.

Bond and Dykstra recognized the virtues of taking the class as unit of analysis, saying flatly that the class mean is the correct unit for their analyses of treatment effects (not distinguishing this from the half-class). Like other investigators of the period, they overlooked the concept of class-level ATI. Because they saw analysis at the class level as a controversial procedure, they compared estimates of treatment effects from their half-class-level analysis within projects and their pupil-level analysis. They pointed out that differences in procedure (especially covarying out Pintner in the individual analysis and blocking on it in the half-class analysis) obscure the comparison. The mean differences as well as the significance levels were greater in the individual analysis, by factors as large as 8 to 1. Higher significance levels are to be expected, because of the increase in claimed d.f.; the increase in means, however, was not explained (see our Section 8).

Procedures in our reanalysis. Professor Dykstra supplied us a set of data for reanalysis, including only the pupils for whom second- as well as first-grade data were available. Moreover, we discarded classes where many data were missing or punched as zero. A zero punch sometimes implies a missing score; even if that is not the case, numerous zeros imply questionable test administration. Our analyses in the first grade therefore cannot match the original report.

We used data from three projects that compared B and LE. The numbers of students for whom we received data were 211, 189, and 181 for B, and 171, 183, and 199 for LE. We dropped two LE classes with many zero scores from one project, reducing N for that project from 199 to 169 students in 8 classes. Two B classes in that project exhibited many zero scores on pretests other than Pintner. We retained these classes in the analyses using the Pintner pretest, but dropped them in analyses of other pretests, lowering N from 181 to 146 students in 8 classes. Likewise, in analyzing Pattern Copying and Identical Forms we set aside an LE class where zeroes were frequent. After cleaning we had from 8 to 11 classes within a treatment available for analysis.

We formed a composite outcome score (POST) from the Reading and Paragraph Meaning subtests of the Stanford Achievement Test, weighting each inversely with respect to its s.d. within treatments pooled. Here we shall not discuss the remaining Stanford Achievement Test subtests. Our conclusions about the contrasts between levels of analysis of the composite are supported, however, by analyses on Spelling and on Study Skills.

We used the following pretests in the reanalysis: Pintner, Murphy-Durrell Letter Names and Learning Rate, Thurstone Pattern Copying, Thurstone-Jeffrey Identical Forms, and Metropolitan Listening Tests. We did not consider the Phonetics or Word Meaning subtests because of the prevalence of missing data in our sample. For demonstration purposes we take up one predictor at a time here, though a multivariate analysis would be more adequate.

Pretest intercorrelations

[were low. Therefore we calculated only univariate regressions. The composite outcome and all pretest variables were standardized to mean zero and s.d. 100 over all cases pooled. POST thus becomes ZPOST, Pintner becomes ZPINT, etc.

Results of our analysis. We obtained a conventional regression coefficient (cases pooled) within each treatment within each project, across all projects within a treatment, and across all treatments within a project. These are more or less conventional analyses. Second, we have a between-classes regression coefficient within each treatment from the analysis of class means within a project. Third, we have a set of pooled within-classes regression coefficients, each calculated as the mean of specific within-class slopes, for the combinations of treatment and project listed above. Fourth, we have the regression within each class. In order to simplify, the tables to follow report data for only three pretests, but conclusions are generally drawn from six such tests.

(1) Conventional regressions. The conventional regression coefficients of the standardized composite outcome variable (ZPOST) onto the standardized readiness measures appear in Table 5.1. The slopes of ZPOST onto the standardized scores for ZPINT, ZIDEN (Identical Forms), and ZLIST (Listening) were higher in LE than in B for all cases pooled within a treatment. These differences generally reappeared within projects. Differences of around 0.25 are neither dramatic nor trivial. Taking that figure at face value, a student 2 s.d. below the mean on ZPINT will rise 1/2 s.d. in posttest performance if he moves from LE to B.

In the regressions of ZPOST onto ZIDEN, the only large effect appeared in the Stauffer project, where the slope in B was close to zero. In the Stauffer project there were rather large slope differences of the same kind on ZLIST, ZLET, ZLRN, ZCOPY, and ZIDEN. In the other two projects slope differences were usually negligible.

We move on now to decompose the effects.

at 5.18

Table 5.1. Conventional unstandardized regression coefficients of ZPOST for individuals pooled

		Project			
Predictor	Treatment	Cleland	Hahn	Stauffer	Projects pooled
ZPINT	B	.09	.43	.30	.31
		(s.d. = 88.3)	(93.1)	(94.8)	
	LE	.30	.73	.62	.56
		(84.4)	(82.4)	(134.1)	
	Pooled	.17	.57	.52	.45
	Difference	.21	.30	.32	.25
ZIDEN	B	.17	.27	.09	.20
		(101.0)	(89.4)	(118.4)	
	LE	.29	.30	.87	.42
		(69.9)	(108.9)	(76.5)	
	Pooled	.16	.28	.32	.26
	Difference	.12	.03	.78	.22
ZLIST	B	.06	.24	.14	.22
		(99.2)	(89.1)	(106.7)	
	LE	.15	.34	.44	.39
		(78.3)	(93.0)	(116.7)	
	Pooled	.10	.29	.33	.31
	Difference	.09	.10	.30	.17

(2) Between-classes regressions. Between-classes statistics weighted by class size for three variables appear in Table 5.2. The slope differences (LE - B) for all six pretests may be summarized as follows:

	-0.30 to -0.01	0.00 to 0.29	0.30 to 0.59	0.60-
Projects pooled (cf. Table 5.1)		3	2	1
Cleland	2		1	3
Hahn	2	2	1	1
Stauffer			1	5

Many differences that seem practically important appear, all the large differences indicating a steeper slope in LE. That is, in LE the abler classes do conspicuously better than classes of low average readiness. Between-classes analysis -- which to us as to Bond-Dykstra seems to be the appropriate emphasis in this research -- paints a far more emphatic picture of ATI than the conventional analysis. The variation from project to project is noteworthy. The Stauffer classes were, as a set, far below the others on most of the pretests, which may or may not be a causal factor in generating slope differences.

Each of the slopes is determined by 11 or fewer classes, and consequently we can have no confidence that similar results would appear in new samples of classes. As in the Anderson reanalysis, plotting data points is instructive. In the Cleland project, the slope onto ZPINT was negative, suggesting that B was detrimental to high-ability classes. The plot of B class means for that project (Figure 5.4), however, showed that the negative slope resulted from the deviation of just one class (near -100, +100). If that one class were deleted, the slope in B would be positive, not negative.

at 5.19

Table 5.2. Between-classes unstandardized regressgion coefficients of ZPOST

		Project			
Predictor	Treatment	Cleland	Hahn	Stauffer	Projects pooled
ZPINT	B	-0.21	.45	.58	.28
		(s.d. = 55.7)	(34.5)	(25.6)	
	LE	.43	1.08	.97	.70
		(44.5)	(29.8)	(86.1)	
	Pooled	-0.01	0.85	0.98	.56
	Difference	.64	.63	.39	.42
ZIDEN	B	.15	-0.07	-0.12	.29
		(70.3)	(50.0)	(43.8)	
	LE	.79	.16	2.70	
		(28.4)	(68.4)	(33.7)	.68
	Pooled	.14	.06	.50	.32
	Difference	.64	.23	2.82	.39
ZLIST	B	-0.08	-0.07	-0.00	.36
		(45.7)	(40.5)	(26.1)	
	LE	1.17	.40	1.91	1.25
		(21.1)	(33.1)	(36.5)	
	Pooled	.12	.13	1.32	.69
	Difference	1.25	-0.47	1.91	.89

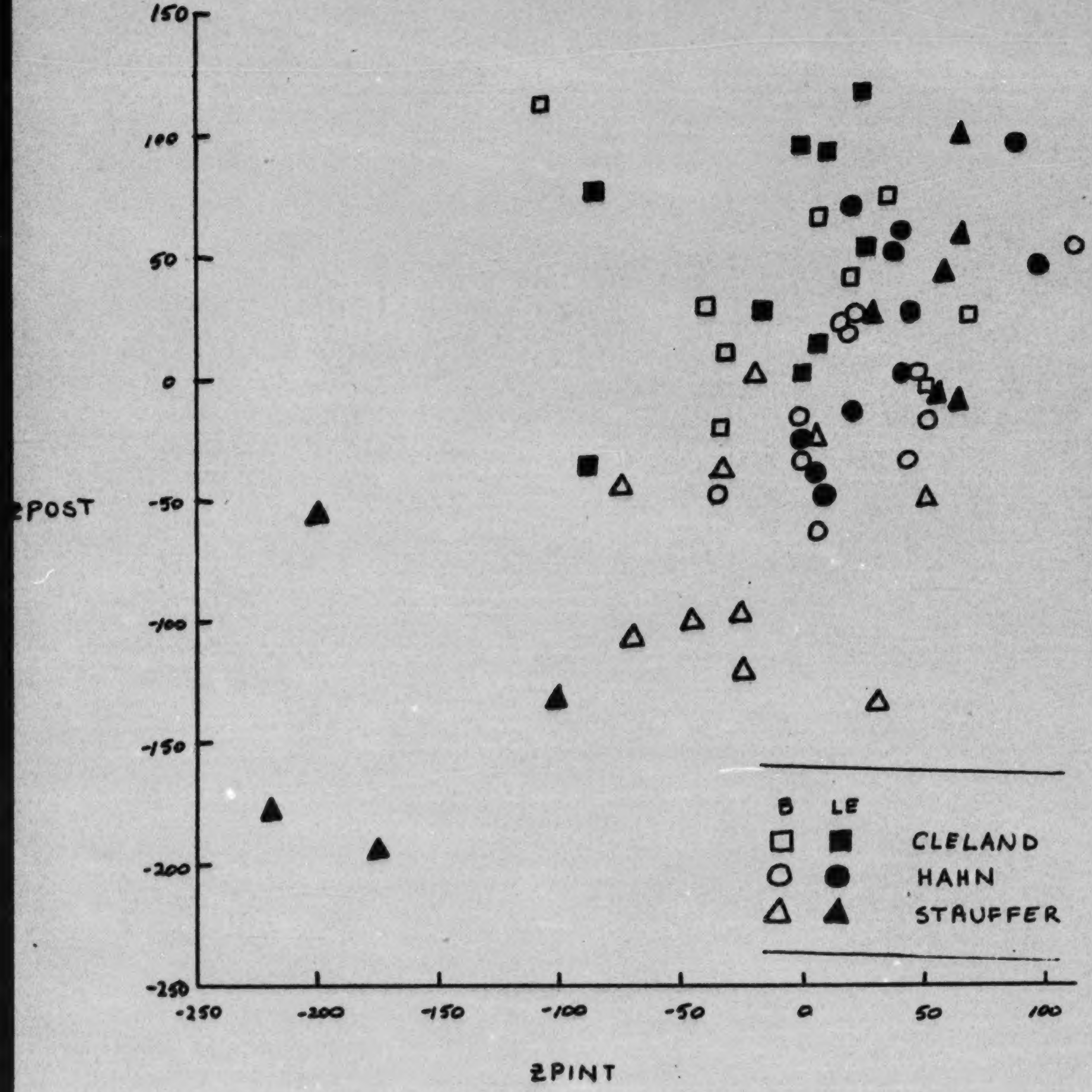


Figure 5.4. Plot of class means in Reading study

The slope differences onto ZPINT in the Hahn project may also be a chance result. The B and LE classes could fit into the same joint distribution. The range on ZPINT is small and the slopes not well determined.

The LE classes in the project directed by Stauffer formed an unusual distribution on ZPINT, split in the middle; some classes were very much lower at pretest and posttest than the majority of classes in that project. In the range where ZPINT > -50, the 8 Stauffer B classes have conspicuously poorer outcomes than his 6 LE classes in that range. This is an effect worth noting, whether significant or not. There is no warrant for a statement about the regression slopes, in view of the narrow range of these 6 LE classes.

The result for ZIDEN agrees with Lo's finding of an interaction of treatment with "perceptual speed". On each subtest -- even ZIDEN -- however, LE and B plots fit into the same joint distribution. Thus, within projects, every negative slope of ZPOST onto ZIDEN or ZLIST resulted from a single class with high posttest and low pretest or from a class with high pretest and low posttest (or from one class of each kind). Again, we must dismiss the differences between slopes among B and LE classes as chance results.

At the between-classes level no Aptitude x Treatment interaction has been established. It is entirely likely that differences in regression slopes are chance results. Studies with many classes per treatment are needed to estimate between-classes effects.

(3) Within-class regressions. The within-class regression coefficients, averaged within each project, appear in Table 5.3. These varied much less than the between-classes slopes, and the differences were small. Just one of the 18 differences exceeded 0.30 (Hahn on ZLIST). It is evident that the interaction effects found in the overall analysis arose from the between-groups differences (which we recognize as likely to be chance effects) and not from within-group differences. If the between-groups effects are untrustworthy, it follows that the differences observed in the overall analysis are untrustworthy.

Taking all projects together, the difference for ZPINT is in the direction of a finding reported by Bond-Dykstra and Lo from their conventional analyses -- but the effect is very weak (0.46 vs. 0.34). On the other pretests the differences for projects pooled are even weaker. Within projects separately, there is not even a consistency in sign between the slope differences in Table 5.2 and those in Table 5.3.

The two comparatively large differences (both in the Hahn project) cannot be taken seriously.

Regression slopes for ZPOST onto ZPINT within Hahn's 11 classes ranged from 0.23 to 0.80 in B and from 0.46 to 1.00 in LE. The slopes of ZPOST onto ZLIST ranged from -0.70 to 0.81 in B and from 0.08 to 0.47 in LE. This undermines the conclusion that slopes tend to be higher in LE than in B.

We looked further into instances where a specific within-class coefficient on a pretest was negative. All these negative values were traceable to one or more anomalous students who scored high at pretest and very low at posttest or vice versa.

at 5.21

Table 5.3. Unstandardized regression coefficients of ZPOST for individuals within classes

		Project			
Predictor	Treatment	Cleland	Hahn	Stauffer	Projects pooled
ZPINT	B	.24	.42	.36	.34
		(68.5)	(86.5)	(91.2)	
	LE	.23	.70	.43	.46
		(71.7)	(76.8)	(102.8)	
	Pooled	.23	.56	.39	.40
	Difference	-.01	.28	.07	.12
ZIDEN	B	.16	.43	.36	.31
		(72.5)	(74.1)	(110.0)	
	LE	.20	.42	.25	.28
		(63.8)	(84.7)	(68.7)	
	Pooled	.18	.42	.27	.30
	Difference	.04	-0.01	-0.11	-0.03
ZLIST	B	.12	.31	.24	.22
		(88.0)	(79.4)	(103.5)	
	LE	.10	.72	.15	.33
		(75.4)	(86.9)	(110.8)	
	Pooled	.11	.51	.19	.28
	Difference	-0.02	.41	-0.09	.11

Conclusions regarding units of analysis. The reanalyses we have made of the Bond-Dykstra data -- of which only a fraction appear in this report -- demonstrate the central themes of our theoretical section.

1. Analyses of the conventional kind, pooling individuals across classes, combine between-class and within-class effects in the sample. They therefore give an equivocal descriptive picture of the relation of outcomes to predictors. We shall later see (Section 8) that they give a poor description of adjusted treatment effects. Significance tests based on individual-level analysis are unacceptable when classes are the unit of sampling. Because between-class data weigh heavily in the overall regression slopes, any undependability in the between-class results casts doubt on the overall results.

2. Between-class analyses appear appropriate in this study. Between-class regression slopes often differ greatly between treatments. These differences, however impressive they may be when coefficients are compared, are evidently dependent on the inclusion of particular "outlier" classes in the sample. With samples of 11 or fewer classes per treatment, observed differences in between-class coefficients are untrustworthy.

The alternative of pooling projects for a between-classes analysis leaves us with modest but consistent differences in coefficients, based on the unusually large sample of about 30 classes per treatment. Whether it is legitimate to combine projects, however, is questionable.

3. Pooled-within-class within-project coefficients do not differ greatly. Even though these coefficients are based on large numbers of observations, their statistical stability is low, because the specific within-class coefficients differ considerably. Such variability may be an important subject for investigation.

Head Start Planned Variation*

Our third set of reanalyses exploits a fraction of the data collected in the Head Start Planned Variation study. This study was carried out in 1969-71, in the wake of the Westinghouse study of Head Start. Like the Follow-Through study that Abt analyzed, this was a prospective study in which a number of sponsors set up experimental classrooms using their own "models" of instruction; the control groups (chosen by the sponsor) were enrolled in "regular Head Start classrooms". Emphasis, however, was to be placed on the contrast among experimental groups. The samples given the various treatments were not chosen to be similar at the outset.

My interest in these data was aroused by an ATI study made by the Huron Institute (Featherstone, 1973) under the direction of Marshall Smith. A number of interactions of treatment differences with such variables as the Pre-School Inventory of Caldwell (PSI) and prior preschool experience (PPE) were reported. Featherstone analyzed data from two cohorts. The 1969-1970 data were used to identify hypotheses for more formal testing on the sample of the next year. The first analyses are said to have been made by "the Data-Text packaged program for unweighted-means analysis of covariance." Huron argued that the variation in sample size from model to model was fortuitous, leaving no reason for giving greater weight to models which had more children. Although this reasoning appeals to me (models being considered fixed), classes are random within models and should be weighted by size within models. It appears that the child was taken as the unit of analysis in both the first and second set of data, and I do not know how the computer package resolved the weighting problem.

---

*Lynne Gray assisted in this analysis.

For analysis of the 1970-71 data, Smith set out a most unusual set of procedures (summarized in Featherstone's appendix). The description is too limited to remove ambiguity; just what the Huron group did is not greatly important here, however, as I am not retracing their footsteps. Some comment does seem to be called for. Let me consider their "PSI regression 4b" (Featherstone, p. 188). The dependent variable was the PSI posttest (PSI2) and the independent variable was "directiveness of model". For this purpose, all Engelmann-Becker cases and all Bushell cases were coded as more directive; and EDC, Bank Street, and Far West cases were coded as less directive. Only 183 cases were employed. Featherstone speaks of 12 first-order predictors (one being model-group and the others being descriptive of the child and his background). Class identification was ignored. The full model also contained 32 first-order interaction terms (11 of them being relations of the form Model x Child characteristic), and at least four second-order interactions but possibly a much larger number. We are told that "regressions were done stepwise with main effects forced in and interactions allowed to enter one by one to explain the maximum additional variance. Results given in the text are for the step on which the standard deviation of the residuals was minimum."

¶ The "results" take the form, first, of the standardized regression coefficient and its significance for each of three variables: Model group, PSI (pretest) and their product. The latter two were significant.

Second, there is a table of "Effects on adjusted PSI [2] score (given in s.d.'s)" at p. 111 -- a fourfold table, crossing directiveness with High/Low PSI-1. The High/Low contrast gives means 0.5/-0.8 in the less directive treatment and 0.4/-0.1 in the directive treatment; the student of ATI effects could easily believe that the choice of treatment for Lows makes a large difference.

Attempts by various persons at Huron to provide me with a more complete description of the analytic procedure broke down, and I can only try to invent a plausible way to get such figures. Perhaps Huron tested significance of the three contributions independently, by the step-down method of removing each one from the full model. Possibly the adjusted scores were deviations from estimates obtained for individuals using the full-model regression equation less the critical terms for PSI-1, Directiveness, and PSI x Directiveness. A procedure even approximately like this would be enormously daring, since it seems to abandon entirely the customary assumption of homogeneous regressions across treatments, and fits dozens of regression coefficients to obtain an adjustment. The final variable is not PSI-1; it is PSI-1 with dozens of things partialled out. Such steps could be given a strong justification, provided that (1) the variables on which treatment groups differed at pretest are highly reliable; (2) the product terms were formed by multiplying deviations from the grand mean -- anything else allows correlations among predictors to totally obscure what is happening-- and (3) children had been sampled and treated independently. I suspect that all these requirements were violated, but it is the third that brings me back to the point of this report.

When there are 183 cases and dozens of correlated predictors, any one of the partial regression coefficients is likely to be highly unstable. This is true even when no one intercorrelation of predictors is large. What is worse, in this study the students were treated in groups. It seems that data come from some 15 classes. (Children for whom IQs were missing were left out of Featherstone's Regression 4). The number of classes is much less than the number of variables entering the regression equation. If classes are the sampling unit, no degrees of freedom are left for making estimates of effects. The data have been seriously "overfitted"; it is not unlikely that the final regression weights in the full model were fitted to rounding error. Analysis at the individual level can be defended, I think, only by asserting that each child received independent treatment, and that differences in pretest characteristics and treatment delivery, among children sampled one from each class, were no larger than would be found for a random sample of children within a class.

We made simple regression analyses, one the conventional overall analysis such as Featherstone employed, and two with partitioned effects. (The analysis we made and the ancova to appear in Section 8 are more nearly like Featherstone's "regression 3" than the analysis just reviewed. The reason for reviewing analysis 4b is that Featherstone's description of it is less equivocal than that for 3; moreover, the only summary data reported on her PSI studies came from 4b.)

In a file of data supplied by Tony Bryk of Huron, we selected a set of 244 children in 18 classes of the more directive programs (Bushell, Engelmann-Becker) and 315 in 30 classes of the less directive programs (EDC, Bank Street, Far West), to investigate the regression of the PSI posttest (PSI-2) on PSI pretest (PSI-1). Featherstone used 422 children in her regression.

As in the Anderson and Bond reanalyses, we have regression slopes within treatments for three analyses. The raw-score means of PSI-1 were 38.6 and 35.1 for the directive and nondirective groups (hereafter D and ND, respectively). The means of PSI-2 were 49.89 and 44.78. The s.d.'s were in the range 9-13. We converted all variables to a metric with 100 as the s.d. for all cases together.

The three analyses all indicate a steeper slope in the ND treatment (Table 5.4). In the Featherstone report also, the D treatment appeared to be advantageous for children low on the PSI pretest and not for Highs. The slope difference in the conventional analysis (our counterpart of Featherstone's) is considerably smaller than the difference in the between-groups coefficients, however. Taking the coefficients at face value, the between-groups value seems to imply that what happened in the ND treatment depended strongly on the ability level of the class; this was much less true in D. Within classes the interaction is considerably smaller than in the between-treatments analysis or the conventional analysis. A reader of Featherstone's report would be led to think that the ND treatment is more profitable for individuals with higher initial PSI, but the within-classes effect is evidently slight. The effect she reported operates mostly between classes. If this phenomenon were established as stable, it would argue that ND is an advantageous treatment for the child placed in a group with high average PSI-1; this would be true whether he himself is high or not.

at 5.26

Table 5.4. Regression coefficients (PSI-2 on PSI-1) within Head Start treatments calculated by various procedures

	η_X^2	Regression coefficients		
Treatment		Conventional	Between classes	Children pooled within classes
Directive	0.34	0.617	0.621	0.615
Non-directive	0.38	0.869	1.083	0.737

Table 5.5. Regression coefficients of PSI-2 on various predictors

		Regression coefficients		
Predictor	Treatment	Conventional	Between	Pooled within
Age	D	0.33	0.39	0.14
	ND	0.52	0.63	0.20
Prior	D	0.26	0.94	0.14
preschool	ND	0.07	0.22	-0.02
White	D	0.14	-0.20	0.27
(v. Black)	ND	0.39	0.57	0.19
MOMED	D	0.21	0.11	0.24
	ND	0.08	-0.37	0.21

The truth of the matter once again is found in a plot of class means. The statistics and the plot are given in raw-score units.

[The statistics did not suggest a dramatic disparity. On PSI-1, there was a mean and s.d. of 38.6 and 7.0 for D, compared with 35.1 and 7.9 for ND. On PSI-2 the values were 49.9, 4.7; 44.8, 7.8. Any drama has to be found in the s.d.'s for PSI-2. It turns out that a small army of ND classes had means below 30 on PSI-1, whereas only three D classes were so low. The unimpressive one-point difference in pretest s.d.'s represents variation impressive to the eye in the chart. Figure 5.5 repeats the story of earlier plots in this chapter: the two sets of points are close to indistinguishable, so far as trend is concerned [save for one lone outlier]. It would be imprudent to assert that there is nothing to the view that the D treatment is comparatively likely to produce changes in class rankings. But the evidence for an interaction is much less impressive than a total of 559 cases and an overall slope difference of 0.25 led us to expect.

Analyses of other variables give further examples of contrast among regression coefficients of the three types. In Table 5.5 are exhibited relations of PSI-2 to various predictors, again with 100 as the overall s.d. for all variables. With regard to age, Featherstone (p. 136) reported that the directive models favored younger children in these data. She gave no numerical results to support the statement. Table 5.5 shows a weak tendency toward a flatter slope in D, which is consistent with her statement, but not impressively so. Again, the difference arises mostly from the sketchily determined between-classes slopes.

Preschool experience did not have a statistically significant effect on PSI-2, Featherstone said, but there was a strong effect on posttest IQ.

at 5.27

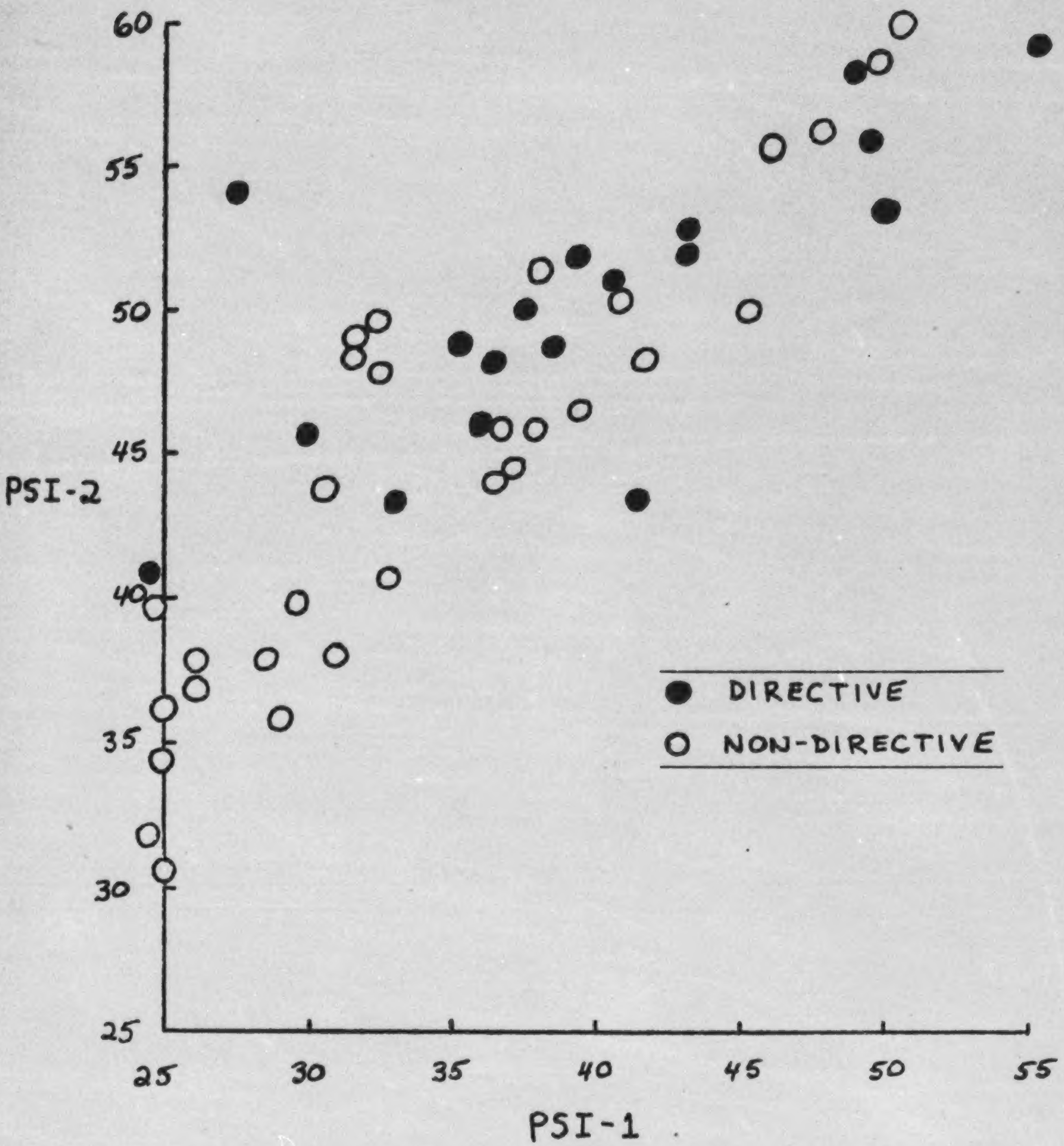


Figure 5.5. Pretest and posttest means on the Preschool Inventory

The estimated regression slope of IQ on preschool experience (after her complex adjustment) was flat in the less directive treatments, and quite steep in D. Since PSI and IQ were strongly correlated, the relations for the two ought to be similar. When we regressed PSI-2 onto preschool experience (with no adjustment) the regression slope was flat with ND in the overall analysis; the slope with D was only modestly positive. The correlation of Age with PSI-1 between classes (0.70) was about as strong as that with PSI-2 (0.66). Very likely Featherstone's complex analysis did not succeed in "correcting" for initial differences in ability. Perhaps her finding arose chiefly from the between-groups slope of ability on age at pretest. Adjusting by means of the shallow conventional slope would by no means remove the large between-classes trend. (See Section 8 on adjustments.)

Featherstone's report on interactions of race is on a within-projects basis. "Three of the less-directive models ... show effects favoring white children, while...[one of] the more directive model[s] shows a highly significant PSI effect favoring Black children" (p. 149). A project-by-project analysis has advantages over a classes-within-treatment analysis. The N's within projects are often too small to allow solid analyses, however. For what it is worth, the breakdown in Table 5.5 shows that if there is a difference between D and ND it is found in the between-classes regression slopes, with the exceptional (but weak and not-to-be-trusted) finding that D classes with more black members earn higher scores on the average than D classes with fewer black members. Again here, the paradoxical reversal of sign was present at pretest.

MOMED was one of three SES indicators in the Huron analysis, and no clear results emerged. The regression slopes in the table show no strong effect and, as usual, the largest difference is found in the weakly measured between-class coefficients. It is hard to credit a finding that classes where the children have, on average, more educated mothers should do less well, as they do in ND. It will come as no surprise that almost the same difference in between-classes coefficients was present at pretest.

Like the Bond-Dykstra study, this is a comparatively large experiment, planned with substantial national resources and subjected to thoughtful attention by both substantive specialists and methodologists over a period of years. Despite the ambitious plan, the study is manifestly too small to permit convincing comparison of the "planned variations", with 500-600 children distributed over eight models and each model represented by fewer than a dozen classes, ill-matched across studies. The Huron analysts had some justification for collapsing so as to contrast the D and ND types of model. They mistakenly thought that with 200-300 cases per treatment they could perform an elaborate search for interactions. In fact, they had 18 cases for most analyses in the D treatment, since a class constitutes a case -- or so this report argues. As has been seen in other studies, the interaction effects reported arise primarily at the between-classes level. Between-classes effects with a limited number of classes are to be viewed with suspicion. Moreover, the effects reported by Featherstone seem with suspicious frequency to be an echo of between-class trends present before training began. The idiosyncracy of the Huron analysis should not obscure the fact that adjustment of posttest scores for initial differences, on the basis of the overall regression coefficient, cannot adjust adequately for between-class differences. I shall return to this topic.

6. Disattenuating regression slopes

In theoretically-oriented work, relations of true or universe scores are the chief concern. In practically-oriented work also, the observed relations ought almost always to be corrected for error of measurement.

In the classical formulation of the problem, ξ_p is the true score of Person p on the test whose observed score is X_p, and ζ_p is the true score underlying Y_p. Then $\beta_{\zeta\xi} = \beta_{YX}/\rho^2_{\xi X}$, the denominator being the reliability coefficient for X. The error in Y does not enter into the correction. Since $\rho^2_{\xi X} \leq 1$, the corrected slope is steeper than the uncorrected slope.

There is no reason to expect the pooled-within-groups and between-groups reliabilities for X to be the same.

The two reliabilities for X will be the same (in the population) if groups are formed at random. Some writers appear to expect between-groups reliabilities to be higher just because means are determined more accurately than individual scores. An example appears in the Abt report (p. V-6); it expresses concern for the biases arising from measurement error in individual-level ancova and then adds: "Measurement error, on the other hand, need not concern us at the school level: the stability of school means is much better than that of the individual child measurements." I shall argue that the standard error of measurement for means is low but that the reliability coefficient need not be higher between than within groups.

It is to be noted that if the within-groups and between-groups regression coefficients are the same, they will not be the same after correction for attenuation -- if the reliability coefficients differ. Conversely, regression coefficients that differ may become equal or may differ in the opposite direction, if the reliabilities differ. This is one more barrier to inferring context effects from a difference in observed regression slopes.

Within- and between-groups reliabilities

Three cases ought to be distinguished:

I. Classes are formed without reference to the scores X. Individuals are tested on X before classes are formed, or persons within a class are tested independently.

II. Class membership is determined in whole or in part on the basis of X. Individuals are tested on X before classes are formed.

III. Classes are formed without reference to the scores X. X measures are taken within the classes, by a group testing procedure.

In any of these cases a variable correlated with X but observed independently of it may influence assignment to classes; i.e., η_X^2 need not be zero in any case. The cases have not been considered separately by previous writers. They require different psychometric analyses.

Let us assume that groups are of uniform size n_c. Also let us assume that all members of a group are tested under the same n_i conditions of facet i (e.g., test items) and the same n_j conditions of j (e.g., occasions). The terminology of this section and the basic concepts come from Cronbach *et al.*, 1972 (hereafter referred to as CGNR). Generalizability theory departs from classical theory in recognizing several sources of error and in not requiring homogeneity of means, variances, and correlations of scores obtained under different conditions. Each person has a universe score μ_p that is the expected value of his scores when X_{pij} is observed on all i, j combinations in the universe. I assume that the same universe is pertinent to every group. I write μ_c for the mean of the universe scores for the group members. Here I write X_p and X_c for the observed scores of individual and group.

We wish to consider at various points parameters of the overall, between-groups, and within-groups distributions. The respective variances -- for example -- will be identified $\sigma_t^2(X_p)$, $\sigma_b^2(X_c)$, and $\sigma_w^2(X_p)[=\sigma_t^2(X_p-X_c)]$.

Case I. When measurement is independent of group membership and vice versa, the generalizability (reliability) of individual-level scores is evaluated without regard to groups, in the manner set forth in OGNR. We can, in the crossed design assumed above, express the observed score as the sum of the universe score, an error component, and a constant: $X_p = \mu_p + \delta_p +$ constant.[1] Estimates of $\sigma_t^2(\mu_p)$ and $E\sigma_t^2(\delta)$ are available for whatever design was used to collect the X scores, and these together provide the coefficient of generalizability $E\rho_t^2$. It is necessary to speak of expected values of certain variances and coefficients because the CGNR model does not assume uniformity of error variances; the conditions i and j drawn for a particular realization of the measuring operation produce a certain population variance for δ or X , and it is the expectation over i and j that is of interest.

For purposes of disattenuating a between-groups regression coefficient, however, one wants a group-level coefficient of generalizability. This is the ratio of $\sigma_b^2(\mu_c)$ to $E\sigma_b^2(X_c)$. The basis for forming groups determines an intraclass correlation $\eta^2(\mu_c , \mu_p)$, or simply $\eta^2(p)$. That is the ratio of between-groups variance in μ_p to the total variance. Since, when p is a member of c , $\mu_p = (\mu_c) + (\mu_p - \mu_c)$,

$$\sigma^2(\mu_c) = \eta^2(p)\sigma_t^2(\mu_p) \quad \text{and} \quad \sigma^2(\mu_p-\mu_c) = [1-\eta^2(p)]\sigma_t^2(\mu_p) .$$

Also, $\delta_p = \delta_c + (\delta_p - \delta_c)$, where δ_c is the average of δ_p over members of the group. Since classes are formed without regard to δ , $E\sigma_b^2(\delta_c) = \frac{1}{n_c} E\sigma_t^2(\delta_p)$ and $E\sigma_w^2(\delta_p - \delta_c) = \frac{n_c-1}{n_c} E\sigma_t^2(\delta_p)$. Even though persons are fixed, the errors of measurement are random. E refers to the expectation over repeated applications of the same design. The between-groups coefficient of generalizability r_b is

$$r_b = E\rho^2(\mu_c , X_c) = \eta^2(p)\sigma_t^2(\mu_p) / [\eta^2(p)\sigma_t^2(\mu_p) + \frac{1}{n_c} E\sigma_t^2(\delta)] .$$

This can be estimated from a G study on individuals which estimates $\sigma_t^2(\mu_p)$ and $\sigma_t^2(\delta)$, and from the observed $\sigma_b^2(X_c)$. The formulas for such a G study are given by CGNR.

No 1 An alternative is simply to carry out a G study on class means.

The within-groups coefficient r_w is

$$E\rho^2 (\mu_p - \mu_c , X_p - X_c) = [1-\eta^2(p)]\sigma_t^2(\mu_p) \Big/ \Big[[1-\eta^2(p)]\sigma_t^2(\mu_p) + \frac{n_c - 1}{n_c} E\sigma_t^2(\delta)\Big].$$

This is a coefficient for classes pooled. Since $\eta^2(p) \geq \frac{1}{n_c}$, $r_b \geq r_w$ in Case I.

If r_t indicates the overall coefficient,

$$1 - r_b = \frac{(1-r_t)}{n_c \eta^2(X)}$$

$\eta^2(X)$ is less than $\eta^2(p)$ unless $r_t = 1$. It is likely that $n_c \eta^2(X) > 1$, hence $r_b > r_t$.

The between-groups coefficient derived here is the same as that suggested by Shaycoft (1962), for which Haney (1974a) presents a derivation. Students are treated as fixed within classes. Haney goes on to discuss an alternative offered by Wiley (in Wittrock & Wiley, 1970) which treats students as random within classes.

Wiley seems to contemplate that a group (the student body in one school, or the class assigned to one teacher) could have many "parallel forms" drawn in the same manner but not randomly representative of the total pool. His question is how strongly class means would correlate from one such set of classes to their set of Doppelgängers.

Case II. When assignment is based in part on the observed X, it is necessary to consider not only $\eta^2(p)$ but also an intraclass correlation $\eta^2(\delta_p)$ or $\eta^2(\delta)$. If assignment takes into account X and at least one other variable correlated with μ_p, $\eta^2(p) > \eta^2(\delta)$. If assignment takes into account X plus variables uncorrelated with μ_p, $\eta^2(p) = \eta^2(\delta)$. Now $E\sigma^2(\delta_c) = \eta^2(\delta) E\sigma_t^2(\delta)$. (In the limit as $\eta^2(\delta)$ decreases, this degenerates to $\frac{1}{n_c} E\sigma_t^2(\delta)$; i.e., to Case I.) With persons fixed within groups,

$$r_b = E\rho^2(\mu_c, X_c) = \eta^2(p)\,\sigma_t^2(\mu_p)/[\eta^2(p)\,\sigma_t^2(\mu_p) + \eta^2(\delta)E\sigma_t^2(\delta)]$$

$$r_w = E\rho^2(\mu_p - \mu_c, X_{p_c} - X_c) =$$

$$\left[1-\eta^2(p)\right] \sigma_t^2(\mu_p)/\left[\left[1-\eta^2(p)\right]\sigma_t^2(\mu_p) + \left[1 - \eta^2(\delta)\right]E\sigma_t^2(\delta)\right]$$

This between-groups coefficient -- which has not been described in the earlier literature -- is smaller than that for Case I. This within-groups coefficient is larger

Case III.[2] Analyses in Chapter VII of CGNR treat data collected in groups, but do not consider simultaneously the generalizability (reliability) of individual and group data. In Case III it is necessary to analyze somewhat differently than in Cases I and II.

Here again, assume persons fixed within groups $(p : c)$.[3] The universe consists of test forms i crossed with occasions j. The investigator intends to generalize from D-study data generated by applying the same form to all groups, each group being observed on one occasion. Such a study has the design $[(p \times j):c] \times i$, $n_i = 1$, $n_j = 1$.

In the G-study, however, each of the k groups is to take more than one form. Let us suppose that each group takes the same n_i forms, each form on a different occasion. The design, then, is $(p:c) \times i$; (j,ci) . The observed score for group c on any one i,j pair resolves into components in this way:

$$X_{cij} = \mu + (\mu_c - \mu) + (\mu_i - \mu)$$
$$+ (\mu_{cij} - \mu_c - \mu_i + \mu) + e_{cij}$$

Analysis of variance produces Table 6.1. Variance components are estimated by entering the actual mean squares in place of the EMS. I write $\sigma^2(c)$ for $\sigma^2(\mu_c)$, etc., as in CGNR.

The between-groups reliability coefficient r_b is given by the ratio of $\sigma^2(c)$ to $E\sigma^2(X_{cij})$, where the latter is defined by the D-study design. With the design specified (one form, one occasion),

$$E\sigma^2(X_{cij}) = \sigma^2(c) + \sigma^2(ci) + \sigma^2(res)$$

In view of the specification that persons are fixed within groups, and in view of the intent to adjust a pooled-within-groups regression slope, it is appropriate to decompose $X_{p_c ij} - X_{cij}$. The components can be written

$$X_{p_c ij} - X_{cij} = (\mu_{p_c} - \mu_c)$$
$$+ (\mu_{p_c ij} - \mu_{p_c} + \mu_c) + (e_{p_c ij} - e_{cij})$$

The last two components are confounded in the G-study design. The analysis of the deviation scores gives the quantities in Table 6.2. (The analysis of variance could be carried out in one step for both

at 6.6

Table 6.1 Mean squares in the analysis of group means and equations for expected mean squares

Source of variance	d.f.	Mean sq	Expected mean square as a function of variance components
Groups c	$k - 1$	MSc	EMSc $= \sigma^2(\text{res}) + n_i\sigma^2(c)$
Forms i	$n_i - 1$	MSi	EMSi $= \sigma^2(\text{res}) + k\ \sigma^2(i)$
Residual	$(k - 1)(n_i - 1)$	MSres	EMSres $= \sigma^2(\text{res})$

Table 6.2 Mean squares in the analysis of individual deviation scores and equations for expected mean squares

Source of variance	d.f.	Mean sq	Expected mean square as a function of variance components
Person within class p_c	$k(n_c - 1)$	MSp:c	$\sigma^2(\text{res:c}) + n_i n_j \sigma^2(\text{p:c})$
Residual	$k(n_c - 1)(n_i - 1)$	MSres:c	EMSres:c $= \sigma^2(p_c i, p_c j, p_c ij, e) = \sigma^2(\text{res:c})$

individuals and groups. It would also be possible to make the analysis shown in Table 6.2 for one class at a time.)

These equations apply to the D-study:

$$E\sigma^2(X_{p_c ij} - X_{cij}) = \sigma^2(p:c) + E\sigma^2(res:c)$$

$$r_w = E\rho^2(\mu_{p_c ij} - \mu_{cij},\ X_{p_c ij} - X_{cij}) = \sigma^2(p:c)/E\sigma^2(X_{p_c ij} - X_{cij})$$

Compare this with r_b. The numerators are $\sigma_b^2(c)$ and $\sigma_w^2(p:c)$. As before, $\sigma^2(c) = \eta^2(p)\sigma_t^2(\mu_p)$ and $\sigma^2(p:c) = [1 - \eta^2(p)]\,\sigma_t^2(\mu_p)$. Ruling out stratified random assignment, the minimum of $\eta^2(p)$ is that for groups formed at random.

[Then $\eta^2(p) \geq \frac{1}{n_c}$ and $\sigma^2(c) \geq \sigma^2(p:c)/(n_c - 1)$.

The denominators of r_b and r_w expand into

$$\sigma^2(c) + \sigma^2(ci) + \sigma^2(j) + \sigma^2(cj) + \sigma^2(ij) + \sigma^2(cij,e) \quad \text{and}$$

$$\sigma^2(p:c) + \sigma^2(p_c i) \qquad + \sigma^2(p_c j) \qquad + \sigma^2(p_c ij,e)$$

Two terms in the upper row have no lower-row counterparts. Within paired terms, the upper term and lower terms are in the ratio $\eta^2/(1-\eta^2)$, where the η^2 is the value for that component. Whether r_b and r_w are near equal depends on the intraclass correlations. Large occasion effects and large intraclass correlations for the pi, pj, or (less likely) pij components will tend to make the group-level coefficient smaller (!) than the within-groups coefficient. Any effect associated with the occasion (noise outside the test room, faulty instructions, etc.) is common to all members of the group. Since these

components of error are not independent over persons, averaging within the group does not necessarily reduce them.
If $\eta^2(p)$ is large and the η^2 for the other three components in the within-groups denominator are small, the ratio of between-groups numerator to within-groups numerator will perhaps be larger than the ratio of denominators. Then the between-groups coefficient is the larger one.

Case III analyses can of course be made for many other experimental designs. The formulas remain to be worked out according to the principles exhibited in CGNR.

An overall "individual-level" coefficient can be calculated by adding the two numerators, adding the two denominators, and then dividing. This is not the value that would be estimated for the overall coefficient by an analysis that ignored groups.

General remarks

I have replaced the "individual" (overall) and "group" reliabilities of other writers with "group" and "individual-within-group" reliabilities. Also, I have separated Cases I, II, and III, whereas other writers confined themselves to Case I without realizing it. The six coefficients will vary in size, but whether the differences are large only future experience can tell us. Surely no one will question the advisability of choosing the logically correct coefficient in any disattenuation procedure.

It should be noted, however, that I have discussed formulas for the coefficient of generalizability only because disattenuation requires a coefficient. In the commonplace investigation of measurement error, the standard error $\sigma(\delta)$ is of far more interest than the coefficient. For most purposes, it is more important to know how well a group is measured than to know whether the measure discriminates between groups. The standard error of generalization of the group mean $[\sigma(X_c - \mu_c)]$ is likely to be considerably smaller than the within-group standard error $[\sigma(X_{p_c ij} - \mu_{p_c})]$ -- so long as persons are fixed within groups.

Notes for Section 6

p. 6.3 [1] The error δ_p is defined according to the experimental design. The data providing a coefficient of generalizability may not be the same as those used to calculate the Y-on-X regression. Indeed, the G study may be carried out under one case and the D study (the regression study) under another. That may still permit one to determine an appropriate coefficient of generalizability, provided that the groups used in the G study are a sample from the population of groups in the D study. In this report I shall assume that the G-study and D-study data are collected by the same experimental design, on samples formed under the same case.

p. 6.5 [2] The Bowers data (p. 2.) appear to me to be an example of Case III. Hauser has pointed out the importance of correcting regression coefficients for attenuation in reaching a decision about the apparent context effect in the Bowers data, but Hauser assumed that the within-colleges coefficient would be small compared to that between-colleges. In Bowers' study, a mail questionnaire on attitude and behavior went to students at many colleges. The only facet along which it seems reasonable to classify individual data is occasions. No doubt variability would appear if the questionnaire had been filled out on two independent occasions. I suspect that there are systematic College x Occasion effects, even if all mailings took place in the same month. A cheating scandal erupting on one campus would cause student responses to a question on cheating to shift, the shift being fairly uniform within that college and not appearing in other colleges. If the question about having been drunk is asked before the main event of the social year on one campus and, by the vagaries of the local calendar, is asked subsequent to that event on another, we can again expect appreciable variability over occasions that is to be considered a group-related error.

p. 6.5 [3] Read: "p nested within c" . The code for designs follows CGNR.

7. Statistical inference

The investigator will wish to generalize formally or informally beyond his sample. In the problems considered here, it seems to me that statistical inference should center on setting confidence intervals on parameters within one treatment. I prefer confidence intervals (or posterior distributions) to tests of the null hypothesis for many reasons, the most compelling one being that in research such as we are discussing the null hypothesis has a high probability of survival. Confidence intervals enable one to report what he found with due caution, yet without suggesting that his study added nothing to knowledge. Posterior distributions have the added advantage that, in principle, they enable experience to accumulate whereas other procedures treat each study as a new venture.

I propose to discuss only limited aspects of statistical inference. Within one treatment, we have to think about the between-groups and within-groups regressions (common and specific) of Y on X. I assume that groups are randomly sampled from a population of groups, and that the distribution of μ_{X_c}, μ_{Y_c} is bivariate normal. I assume members fixed within collectives.

I make no attempt to set limits on the regression of Y onto ξ. In time, procedures for setting limits on disattenuated regressions will be wanted.

A full examination would next move on to estimates of the treatment effect. A distinction is required between one-population and two-population (or larger) studies, the former being those where groups were assigned randomly to treatment. The case of homogeneous regressions (traditionally assumed) must be considered separate from the case where within-treatment coefficients differ. Again, error of measurement is to be considered; use of the attenuated regressions in covariance adjustment gives a false conclusion. The difficulties of inference about single treatments or treatment comparisons have not been resolved even for the study where individuals are assigned and treated, with none of the complications introduced by grouping (Cronbach *et al*., 1976).

I omit inference about multiple regressions from this report entirely.

Sampling error of a mean

The simplest statistical inference evaluates the population mean on the basis of sample information. If individuals are the unit of sampling and analysis, the sampling error of the mean is estimated by $s(Y)/\sqrt{N}$. If groups are the unit of sampling and analysis, the corresponding formula is $s(\mu_{Y_C})/\sqrt{k}$ where k is the number of groups.

Suppose all groups are of size n; $N = kn$. Then $\sigma^2(\mu_{Y_C}) = \eta_Y^2 \sigma^2(Y)$. In random sampling the intraclass correlation is $1/n$, and the two modes of calculation will lead to very similar conclusions. The conclusions will not be identical, as the t distribution depends on the number of degrees of freedom claimed.

With larger values of η_Y^2 , the sampling error calculated at the group level -- as it should be -- becomes quite a bit larger than the one calculated at the individual level. Consequently, analysis with groups as units generates comparatively wide confidence intervals. When the null hypothesis is valid, the analysis with individuals as units will report a significant effect unduly often.

The between-groups regression

Where groups are randomly sampled from a population and all receive "the same treatment", the well-known procedures for establishing confidence intervals for a regression line apply to the between-groups regression. The parameters of the regression equation in the population are μ_X, μ_Y, and β_b; the value of $\sigma(\mu_{Y_c})|\mu_{X_c}$ is also pertinent. Each sample is characterized by a pair of means, a coefficient b_b, and an $s_{\bar{Y}\cdot\bar{X}}$. Under the assumption of normality, the b_b are distributed normally about β_b, independently of the $s_{\bar{Y}\cdot\bar{X}}$. The expected joint distribution of b_b, s pairs permits one to define an elliptical confidence region in the b, s space outside which the pair β_b, σ is unlikely to fall. From this comes the usual equation which, for a between-groups regression, can be written (Dixon and Massey, 1969, p. 198):

$$(7.1) \quad \hat{\mu}_Y + b_b(X - \hat{\mu}_X) + t_{\frac{1}{2}\alpha} s_{Y\cdot X}\sqrt{\frac{1}{k} + \frac{(X - \hat{\mu}_X)^2}{(k-1)s^2_{\mu_{X_c}}}} < (E_{\mu_{Y_c}})|X_c$$

This describes the lower confidence limit for the regression line. The upper limit is described by the same expression with $t_{1-\frac{1}{2}\alpha}$ replacing $t_{\frac{1}{2}\alpha}$. There are k - 2 d.f. for t, where k is the number of groups. The two equations describe an hyperbola in the X, Y space; the asymptotes of the hyperbola (which pass through the sample mean $\hat{\mu}_X$, $\hat{\mu}_Y$) identify the outer limits of the regression coefficient.

Confidence intervals for group data are likely to be wide, because in most studies the number of groups is small. If $s_{\overline{Y}\cdot\overline{X}}$ is small, however, as happens in some group data, the confidence interval can be satisfactorily narrow in the neighborhood of the X mean.

Samples on the order of 100 classes are required to make b_b a good estimate of β_b . This statement comes as an unpleasant surprise to most research workers, and some find it hard to believe. A simple example may overcome such doubts.

Suppose that $\sigma(\mu_{Y_c}) = \sigma(\mu_{X_c}) = 1$, and that $\rho(\mu_{X_c}, \mu_{Y_c}) = 0.40$.

Then $z = 0.42$; $\sigma_z = 1/\sqrt{k-3}$. If $k = 100$, $\sigma_z = 0.10$ and the 95% limits on sample values of z are 0.22 and 0.62, implying limits of 0.22 and 0.55 on r . Swings over that range, and over the corresponding range of b , could be consequential. Just how large a sample to demand is a matter of judgment, of course.

The within-groups regression

The usual procedure cannot be adapted to establish a confidence interval for the pooled-within-groups regression if individuals are fixed within groups.

The several β_c are independent estimators of $E\beta_c$. If the β_c are assumed to have a normal distribution, it is a simple matter to set confidence limits:

$$\bar{\beta}_c + t_{\frac{1}{2}\alpha} s_{\beta_c} < E\beta_c < \bar{\beta}_c + t_{1-\frac{1}{2}\alpha} s_{\beta_c}$$

The number of degrees of freedom is $k - 1$. These limits can be thought of as describing two within-groups regression lines both of which pass through the point 0,0. (The hyperbola of the between-groups inference degenerates when the mean is given a priori.) Whether these confidence limits will be wide or narrow depends on the spread of the β_c.

With regard to the specific β_c, it is difficult to ask a useful inferential question in the usual circumstances (see pp. 4 .1-8).

8. Analysis of covariance

Analysis of covariance is used to evaluate the difference among outcomes in two or more treatments. An adjustment for initial differences is the crux of the procedure. Even when assignment of individuals or classes to treatments is random, the choice of unit of analysis has some effect on the result. When assignment is not controlled, the initial difference may be large; then the choice of units may greatly affect the adjustment. The standard procedure is to calculate (directly or indirectly) an adjusted outcome score for each person or group. Should the adjustment be derived from the within-groups, the between-groups, or the overall individual level regression coefficient? Many investigators seem routinely to assume that the regression coefficient calculated at the individual level (within treatments) should be used. Among those who recognize more than one possibility, some carry out and report alternative analyses without a clear basis for interpreting them.

In analysis of covariance, a number of difficulties arise even apart from questions regarding units of aggregation. In the best case, one has an experiment with random assignment; then the analysis with any regression equation gives an unbiased estimate of the treatment effect. The statistical model assumes that the covariate and its values are fixed, and this is not generally appropriate in social and educational research. Problems multiply when selection or self-selection determines who enters and completes each treatment. Poor data on initial characteristics -- failure to measure some characteristic for which adjustment should be made, or inaccurate measurement -- bias the estimate of the treatment effect. I shall say no more about these difficulties, though they are pertinent to research on groups.

Recommendations for analysis of covariance have to take into account the design for collecting data. Any of the elaborate designs to which analysis of variance is applied can be extended by adding covariates, since ancova _is_ anova of adjusted scores. It will be sufficient here to consider two designs, and to limit attention to investigations with only two treatments, A and B.

Design 1 is an extension of the two-group experiment (or quasi-experiment). Collectives are nested within treatments, and members are nested within collectives. Collectives are considered to be a random sample of a population of collectives; if assignment to treatment was nonrandom, there is a population for each treatment, defined by the selection rules, explicit or implicit. I have suggested that members be considered fixed within collectives, but some analyses treat members as random.

Design 2 crosses treatments with a blocking factor. This factor may be a higher-order collective. In the Performance Contracting experiment, schools -- the unit to which treatments were assigned -- were nested within school districts, each district in the study having a school in each treatment. The factor may be a potential cause whose main effect is to be removed from the error variance, as when every teacher handles a class in each treatment. The factor may be a characteristic of persons, as when class membership is determined by selecting students within certain IQ ranges. Once collectives are identified within blocks, they may be assigned to treatments randomly or not. Collectives within a treatment are assumed to be experimentally independent.

Design 1. Collectives nested in treatments

It is usual in educational research to choose one set of schools or classes for Treatment A and to choose independently another set for B.

This was the design in the Follow Through study where Abt (Cline et al., 1974) offered analyses of covariance at the pupil, class, and school levels. "Where results are consistent for parallel questions across the three levels of aggregation", they said, "we have enhanced confidence that they represent the true effects." This was a study with nonrandom assignment and the treatment populations differed in initial characteristics. The analyses would be most unlikely to agree with each other even if Abt had used the same variables in each analysis. Each covariate was formed by multiple regression, hence different composites were used to make the three adjustments. It would require an enormous coincidence for adjustments made with different composites and different regression coefficients to be the same.

Alternative adjustments. It will be instructive to consider a detailed list of alternatives, though some of them seem unreasonable a priori. To keep matters simple, suppose that there is one perfectly reliable covariate X, that there are just two levels, and that corresponding regressions have the same coefficient from treatment to treatment. But do not assume that $\beta_b = \beta_w$. Set the mean value of X for all cases at zero.

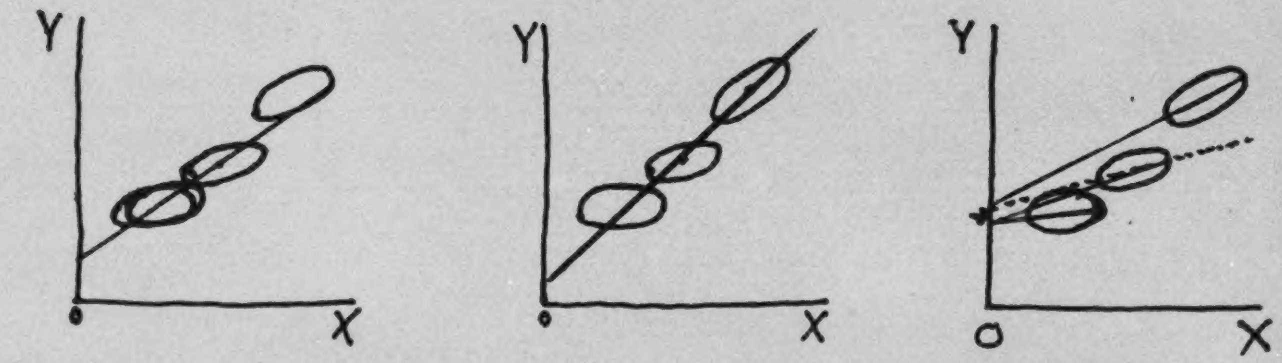
Each analysis determines the intercept of a within-treatment regression line at $X = 0$. The heart of the process is to fix a coefficient $\hat{\beta}$. Then one subtracts $\hat{\beta}X$ from the Y score for each unit of analysis and averages within the treatment. The coefficient may be determined in these ways:

1. Overall. Calculate a within-treatments regression coefficient without regard to boundaries of collectives.
2. Between collectives. Calculate a regression for collective means, within treatments.
3. Within collectives.
    - 3a. Convert scores to deviations from the mean of the collective. Pool collectives, and calculate the regression coefficient. One can obtain an intercept for each collective, or for all collectives.
    - 3b. Calculate a regression equation within each collective, and use it to adjust scores of members. This gives an intercept for the collective.
    - 3c. After calculating within-collective coefficients as in 3b, determine the trend of coefficients as a function of the mean of the collectives on X . For collectives with any X mean, obtain a coefficient on the basis of this regular trend.

In analysis 1, the number of degrees of freedom comes from the number of individuals. In 2, 3b, and 3c, the number of classes is the basis for determining degrees of freedom. In 3a, investigators might adopt either basis for calculating degrees of freedom.

The several analyses are illustrated in Figure 8.1, which shows schematically the data for three collectives in just one treatment. In this illustration the between-groups slope is steeper than any of the within-group slopes, and within-group slopes have a regular trend. Panels (i), (ii), and (iii) represent adjustments 1, 2, and 3, respectively.

at 8.4



(i) Overall

(ii) Between groups

(iii) within groups considering trend

Figure 8.1. Estimates of adjusted treatment mean considering three alternative regression lines

The adjusted mean in Panel (ii) is the smallest of the three. Although it is not necessary that the between-groups coefficient be larger than that within groups, experience shows that this usually is the case. Adjustment by means of the between-groups coefficient, then, is a more drastic adjustment than the others, and leads to a less favorable conclusion about the treatment for which the sample stands higher on X .

The pooled-within-classes adjustment (3a) is shown by the dotted line in Panel (iii); it gives the largest of the adjusted means. The overall adjustment in Panel (i) is close to that in Panel (ii) because of [the large η_X^2 . It will always lie between the adjusted means from the between- and pooled-within adjustments. It will be close to the latter if η_X^2 is small. The adjustment 3c that takes into account the steeper slope in classes with higher X gives the intercept indicated by the arrow in Panel (iii). This is the average of the separately adjusted means for the classes.

If it is believed that group-caused effects are nonnegligible, the between-collectives analysis appears to be appropriate. Each collective is an independent realization of the effects, group-caused and other, sampled from a population of realizations. The intent is to generalize to a population of realizations for which the overall mean X is near zero.

If there are no group-caused effects, then one could still analyze appropriately at the between-collectives level. Collectives are the unit of sampling, and unless collectives differ only by chance on initial variables relevant to Y , the statistical inference has to be at the group level.

I see no way to defend adjustment based on any of the coefficients calculated within collectives. Note that the uppermost regression line in Panel (iii) projects the unadjusted mean for Y in the rightmost collective into a considerably lower adjusted mean. If this has any meaning, it is a prediction as to what would happen if individuals with $X = 0$ were treated as members of a collective with a high mean on X. The extrapolation is rash on its face. To use it in evaluating the treatment, however, embodies the even more absurd extrapolation that this is the mean to be expected "if this class were made up of students whose mean score is zero". The presumed reason for a steep slope in the high-X classes is that the level of X makes a difference in the slope, so that the extrapolation is self-contradictory.

If within-groups slopes are irrelevant, why mention them? My reason is that they crop up in practice! Most obviously, every time an educational investigator performs an experiment with one collective per treatment, his analysis of covariance uses adjustment 3. (Since there is only one class per treatment, cases 3a, 3b, and 3c are indistinguishable, and indistinguishable from the overall adjustment 1.)

The analysis is just, if the class is a random collection of individuals who respond independently to the treatment. If not, the investigator has adjusted without information on β_b. If $\beta_b \neq \beta_w$, he has overadjusted or underadjusted.

In a multigroup study, if one knows or is prepared deliberately to assume $\beta_b = \beta_w$, the overall analysis is justified and the others are less suitable. The overall analysis also makes sense when individuals are sampled individually, and the individual's experience is not systematically associated with that of others in his group. In this latter case, demographic effects may cause a difference between β_b and β_w that has no bearing on the estimate of the treatment effect.

Cooperative Reading data. To illustrate the contrasts among analyses 1, 2, and 3a, Webb and I processed data from the study of Hahn within the Cooperative Reading Program. We used 183 students in 11 classes that had followed a Language Experience (LE) program and 189 in 11 classes in a Basal (B) program. The raw score on Stanford Word Reading at the end of Grade I (a component of POST) was our dependent variable, and the Pintner test of mental ability our covariate. The B group had a Pintner mean higher than the LE group (1.24 points higher; about 0.2 s.d.).

1. The first analysis used the overall regression. Scores within a treatment were pooled without regard to class membership. The overall regression slope for the combined treatments was 0.45. The covariance analysis gave this information:

	before adjustment	after adjustment
Mean for LE	24.5	24.2
Mean for Basal	22.2	22.5
Difference	2.3	1.7

	SS	d.f.	MS	
Treatments	290.94	1	290.94	F = 9.43
Within treatments	11384.02	369[1]	30.85	
Adjustment	4289.71	1		
	15964.67			

To recognize possible homogeneity of regressions we also calculated the coefficient within each treatment separately. The regression coefficients were 0.55 in LE and 0.37 in B, and the corresponding adjusted means were 24.2 and 22.4. The use of the specific coefficients had little effect on the difference in adjusted means, as is to be expected; fitting within a treatment has the effect primarily of reducing the residual variance.

2. When class means were used to calculate the within-treatments regression coefficient, it rose to 0.63. The covariance analysis was carried out as before but with the adjusted class mean as the dependent variable. This score was entered for each class member, in keeping with our policy of weighting. The sums of squares from this analysis were used, but the number of degrees of freedom for the denominator given by the computer, based on individuals, was replaced with 20, based on classes.

	SS	d.f.	MS	F
Treatments	222.25	1	222.25	2.34
Within treatments	1894.01	20	94.70	

The F ratio does not reach significance. An adjusted treatment effect of 1.5 replaces the 1.8 of the individual-level analysis. The adjusted treatment means are 24.1 and 22.6. (Bond and Dykstra reported an adjusted individual-level treatment effect of 1.8, matching ours. After class-level adjustment, they reported an effect of 5.6, however; I have been unable to determine why.)

3. The within-classes coefficient for treatments pooled was 0.42.

The summary in Table 8.1 indicates that the shift in methods of adjustment did not produce a great difference in the adjusted treatment effect. The shift from a claim of significance to non-significance stems from the larger error variance that accompanies the correct number of degrees of freedom.

Follow Through data. One more brief example can be derived from the Follow Through data mentioned in Section 5. Featherstone reported that children with prior preschool experience were better off in a less-directive treatment. The study had classes nested within treatments.

at 8.9

Table 8.1. Coefficients and adjusted means from three analyses

Analysis	Coefficient(s)	Means after adjustment		Difference
		LE	B	
Anova	---	24.5	22.2	2.3
Overall	.45	24.2	22.5	1.7
Between classes	.63	24.1	22.6	1.5
Within classes	.42	24.2	22.5	1.8

Table 8.2. Alternative adjustments of hypothetical data for three collectives within a treatment

Center	Mean SES		Mean MRT		Adjustment		Adjusted B Mean		A - B Difference	
	A	B	A	B	1	2	1	2	1	2
1	7	8	5	2	-3	-1	-1	1	6	4
2	5	6	0	0	-1	-1	-1	-1	1	1
3	3	4	-5	-2	1	-1	-1	-3	-4	-2
All	5	6	0	0	-1	-1	-1	-1	1	1

Adjustment 1 is to overall mean of A's and adjustment 2 is to center mean of A's.

Featherstone found it appropriate to use separate regression lines for the two treatments. As in Section 5, all variables are rescaled so that the s.d. for all cases together is 100.

Here I take the posttest on the Preschool Inventory as dependent variable and preschool experience as predictor. When the same set of data is processed in modes 1, 2, and 3a, the adjusted treatment effects shift as follows:

	ND mean	D mean	Difference
Unadjusted	-0.208	0.269	0.477
1. Adjustment with overall regression	-0.075	0.147	0.222
2. Between-groups adjustment	-0.042	0.147	0.189
3a. Within-groups adjustment	-0.095	0.148	0.243

The between-groups adjustment, which I consider the appropriate one, reduced the treatment effect to about 85 per cent of that reported by the conventional overall analysis, and reduced the numerator of the F ratio by 28 per cent. Change occurred primarily in the value for ND, since b_b and b_w were nearly the same in D.

Design 2. Treatments crossed with blocks; collectives nested

The design in which collectives are blocked gives up some number of degrees of freedom, but brings irrelevant variance under tighter control. Once data are collected under such a design it seems to make no sense to ignore the blocking. The blocks may be regarded as fixed (e.g., when the 50 States serve as blocks) but it is probably more common to regard them as randomly representative of some larger population of blocks. Then the adequacy of the information depends on the number of blocks in the sample.

Suppose that, within blocks, there are just two collectives, one assigned to each treatment. Then, if no covariate is to be considered, one might reasonably form the means for the collectives, take the difference between treatments in each block, and test whether the mean of the differences differs from zero.

An equivalent procedure is a two-way analysis of variance, with the Blocks × Treatment interaction supplying the error variance for the F ratio. Blocking serves the same function as analysis of covariance, insofar as there are relevant initial differences between blocks. Whatever variables contribute to variance in outcome at the block level are extracted; this does not modify the estimate of the treatment effect, but it reduces the estimated sampling error.

A covariate may now be introduced to allow for initial differences between collectives within the block. The question is, how does the mean outcome in the collective relate to the initial mean? And how would the difference in outcome means be altered if the initial difference were zero?

The plan of the Head Start study. My thinking about Design 2 was stimulated primarily by the famous Head Start evaluation made by the Westinghouse Learning Corporation (hereafter WLC; 1969) and the reanalysis by Smith and Bissell (SB; 1970). Both sets of analysts wrestled with the problem of units. Although I shall raise questions regarding the solutions put forth in those reports, WLC and SB were ahead of their time in their thinking about levels of analysis. The entire body of data includes findings for full-year Head Start programs and summer programs, for white children and black children, and for follow-up results on many tests given in Grades 1, 2, and 3. I shall stay within the data processed by Smith and Bissell as well as WLC and shall reduce attention further to the full-year data on children of all races together, with SES as covariate and Total score on the Metropolitan Readiness Test early in Grade 1 (MRT) as outcome. I shall not trace the influence of a subtle shift in covariates that occurred; in one analysis WLC used a single predetermined SES composite and in another formed a three-variable composite post hoc by multiple regression (overall); SB formed three composites, one from the between-group correlations, one within-groups, and one overall (p. 9.18). For my purposes, I shall simply speak of SES as covariate.

Head Start was administered through local offices or centers, each with its own territory; centers were the primary unit of sampling for the study. Within a center there was occasionally more than a single class, but this was too infrequent to be considered in the design. Classes within a center, then, were combined. The pool of Head Start (A) children consisted of those who could be tested in first-grade and who had attended the centers under consideration. A pool of control (B) children consisted of children located in first-grade who came from the center territory, who had been eligible for Head Start, and who had not received this or other prekindergarten training. Now from the A pool a sample of 8 children was drawn for each center; and 8 B children were individually paired with them -- matched on sex, race, and attendance/nonattendance at kindergarten. Thus the design is Treatments crossed with Pairs of children, Pairs nested within Centers, one child per cell. Both WLC and SB, however, ignored the matching at the individual level, and I shall ignore it also. I do not see how the information could be used constructively in assessing the main effect of treatments.

The character of the discriminant may be noted in passing. In this study the collectives differed on several known variables. In addition to the variables in Table 4.1, center pools differed in racial makeup and in the prevalence of kindergarten attendance. This is one of the comparatively uncommon instances of a quasiexperiment in which membership in a collective is correlated with a post-treatment variable, for reasons not arising wholly from initial demographics or from the group as cause. (The child's attendance at kindergarten may have been to some degree a consequence of his Head Start experience, or of his parents' desire to compensate for the absence of Head Start experience.) Whatever the causal chain, if children with kindergarten experience are more numerous in certain centers, and kindergarten raises MRT scores, the discriminant is correlated with $Y{\cdot}X$ on a basis that can neither be considered a context effect nor a demographic effect.

Alternative analyses, assuming homogeneity of regression. WLC reported two analyses, and SB reported six. These eight by no means cover all the possibilities, and I shall argue that none of the eight -- as I understand each of them -- is logically appropriate. It is not precisely clear how some analyses were conducted, and in both reports there are mysterious differences in the unadjusted mean of MRT from one table to another. It is not so important to discuss those particular analyses as it is to comprehend the range of alternatives and to develop a rationale for choice among them. I may as well say at the outset that my preference among the analyses has changed more than once as I have studied the problem, and that I am not convinced that I now know what should have been done with these data.

The usual analysis of covariance adjusts by calculating deviations from the mean for the treatment sample, pooling cases over treatments, and determining the regression coefficient for outcome deviations onto covariate deviations. This assumes that corresponding A and B regression coefficients are the same in the population. SB found this not to be the case at any level of analysis for MRT-on-SES regressions, and so in some analyses they abandoned the homogeneity assumption. I start with analyses that assume homogeneity. It will be helpful to denote the between-collectives regression coefficient as b_b, that within collectives as b_w, and the overall regression which ignores collective boundaries as b_o [2]. I am assuming here that the two b_b agree, that the b_w agree, and that the b_o agree.

The obvious possibilities follow, grouped in three categories. I code them to identify the procedures I think WLC and SB adopted. (E.g., WLC1 resembles one of the two WLC analyses.)

A. Number of individuals as base for d.f.

A-b. Use of b_b. (No one has proposed to use this. I list it for the sake of symmetry.)

A-w. Use of b_w. The product b_w(SES) is subtracted from individual MRT scores. Then to compare treatments an unmatched t-test is carried out on individual adjusted scores. (Where I refer to a t-test, the algorithm of anova or ancova could be used, leading to an F-test.)

A-o. Use of b_o. From MRT is subtracted b_o(SES). This is conventional ancova ignoring the collectives (WLC2). Smith and Bissell made a generalized regression analysis which is similar (SB1). (The significance test is often made incorrectly, by assessing the significance of the increment in mean-square-explained-by-regression when the dummy variable for treatment (T) is added as a last predictor. The treatment effect is

$b_T \sigma_T$ and the variance in Y accounted for by treatment is $b_T^2 \sigma_T^2$, which may be larger or smaller than the increment in mean square. The F ratio is raised or lowered in the same way that overadjustment would distort it.

I see no warrant for taking individuals as the base for degrees of freedom when centers were the primary unit of sampling. Both WLC and SB offer A-o analyses with the idea that they may be "more sensitive" to small differences, but inflating the number of d.f. simply produces spuriously low levels of the α risk.

B. Number of collectives as the base for d.f.

B-b. Use of b_b. Either individual scores or collective means are residualized by subtracting b_b(SES). An unmatched t-test on adjusted means for collectives is run. There is an analogous generalized regression analysis (SB2).

B-w. Use of b_w. As in A-o, but adjusted MRT_c are used in an unmatched t-test.

B-o. Use of b_o. Like A-o, but with an unmatched t-test on adjusted MRT_c.

The salient difficulty here is with the unmatched t-test. Collectives are nested within centers, and no method of adjusting for SES can remove all the relevant differences among centers. Such differences do not falsify the estimate of the treatment effect, but they unnecessarily lower the power of the statistical inference. The obvious way to escape the problem is to use a matched t-test. Whether this is profitable depends on the between-centers variance in adjusted MRT.

C. Number of centers as the base for d.f.

C-b. Use of b_b as in B-b. Matched t-test run on adjusted MRT_c.

C-w. Use of b_w as in A-w and B-w, with matched t-test on means for collectives. WLC did essentially this with analysis of covariance by removing the variance for centers, and then using the Centers × Treatments interaction as an error term to evaluate the Treatment mean square, after covarying out SES on the basis of b_w (WLC1).

C-o. Use of b_o as in A-o and B-o, with matched t-test on adjusted MRT_c.

Which regression makes sense? The use of b_b seems to address the question: If we were to search through a large number of collectives, disregarding center membership, and were to pair up selected A and B collectives so that each pair had the same SES mean, what difference in the MRT_c would be expected? I say this, because using b_b estimates the expected MRT for a collective regarding which one knows the SES mean and nothing else. So this method does not take pairing into account. It does not allow for variables relevant to Y, and orthogonal to X, on which centers differ, so it loses the value of the matching t-test. As for b_o, it suffers as usual from being a composite of b_b and b_w, the b-regression and the w-regression. If b_b is off the mark, so is b_o. Using b_w seems to ask almost the right question. If we search through the pool of A and B children within a center, and select out two sets that have the same SES mean, what mean difference in MRT would be expected? That is, the analysis attempts to simulate the result in an experiment where equivalent children within a center are assigned to treatments. The analysis, however, ignores the fact that A children were treated in a group. If many collectives were formed from the children in the same center and given the Head Start treatment, the between-collectives-within-center-within-A regression need not be the same as the regression within-collectives-within-center-within-A. One might waive the question of demographic differences (other than SES) between such hypothetical classes, but it seems that one must also assume absence of context effects to justify analysis C-o. If this argument is correct, then, one would need more than one A collective within a center to arrive at the logically appropriate adjustment. The problem does not arise with B's, who were treated individually and could not have been subject to a context effect.

This meticulous dissection is required to work toward an understanding of analysis of covariance, but it does not cast serious doubt on the results from WLC1, since the amount of adjustment was small. Insofar as that analysis is in question -- aside from the challenges any quasiexperiment is open to -- the question arises from the inhomogeneity of regressions, which I deal with next.

Alternative analyses, recognizing heterogeneity of regressions. In the analysis with homogeneous regressions, one finds out how far the A cases are above the reference regression line and how far the B cases are below it (or vice versa), combines those differences, and reaches an estimate of the treatment effect. One would have the same result if he formed the two within-treatment regressions (which are parallel) and determined the distance between them. The usual way of speaking about the analysis is to speak of the "intercepts" where the two regression lines cross a reference value of X . Although the choice makes no difference when the regressions are parallel, it makes best sense to think of X as at the mean of the grand population. When the regression lines are not parallel, the treatment effect is different at each value of X. One can compare their intercepts at the grand mean, and many investigators would do just that, to determine the effect of the A - B difference for the population of eligible children. SB chose instead to determine the A - B difference where X is at the mean of the A population, of children who were eligible and who entered. I accept this decision at least for present purposes.

The SB technique was to adjust scores of B children only. The B's were of higher SES than the A's, on average, so their MRT scores were adjusted down, to get an estimate of what the B mean would have been if the B's had had an SES mean comparable to that of A children. If the B regression coefficient is s, and a B child is u SES units above the reference group of B's in SES, then his MRT score is lowered by su units.

An important choice is to be made regarding the reference group. The child is, let us say, in a center where the A mean on SES is 7 (on no matter what scale), whereas the SES mean of all A's is 5. The child, let us say, has an SES of 7; do we take A's in his center as the reference group and make a zero adjustment, or do we take all the A's as the reference group and reduce his MRT score by $2s$? It makes no difference in the final report of the treatment effect, but it does influence the variance.

To show this, let us consider artificial data and assume an analysis with b_b. Let us adjust the MRT_c and not individual scores; the argument would have the same flavor if we adjusted individuals or used another regression coefficient.

The simple data in Table 8.2 show $b_b = 1.0$ for B collectives. In Center 1 the B mean on SES is 8, 3 units above the mean of all A cases; hence adjustment 1 is -3.0. The SES mean of 8 is just one unit above the A mean for that center, and adjustment 2 is -1.0. When the full set of computations is carried out, we see that the mean over centers for B's is the same with either adjustment. The treatment effect changes from 0 to 1 with either adjustment. But the variance of differences, which provides the error term in a matched t-test, is lower when adjustment 2 is used. It seems to me that adjustment 2 is more appropriate than adjustment 1, when matching within centers is intended. (Whether using b_b is appropriate is another question.)

It would be possible to set up categories D, E, F of regressions corresponding to adjustment 1. This would be a pointless digression. Let me say only that as nearly as I can tell SB4, which used b_o, would be coded D-o, as counterpart of A-o, and SB6, which used b_w, would be D-w. With regard to adjustment 2, there is no point in detailing categories G and H, counterparts of A and B; let us consider just the category that has the proper d.f. (counterpart of C).

I. Number of centers as the base for d.f.

- I-b. Use of b_b. Regression of MRT for B collectives on SES_c is determined, and MRT_c for B's in a particular center is adjusted according to the discrepancy of their SES_c from that of A's in the center. Matched t-test is entered with the mean MRT_c of A's and the adjusted MRT_c of B's (SB5).
- I-w. Use of b_w. Similar to I-b save that b_w is calculated for the B's.
- I-o. Use of b_o. Similar to I-b, using b_o.

Once more we face the question, which regression? It seems to me that using b_w is wholly appropriate. B's were treated individually, and they were identified individually within centers. Hence one wishes to estimate the probable MRT score for individual children within a center who have a particular SES mean. I can see no way in which between-center differences among children (which enter the other two regression coefficients) become relevant to a within-center adjustment.

Although analysis I-w, which I favor, came to light because regressions were heterogeneous, I believe it also would be justified if corresponding regressions had happened to be homogeneous. It will be recalled that the regression coefficient required to recognize the grouping of A cases could not be evaluated. The Head Start design, with one treatment given individually and the other treatment given to groups, treatments being crossed with primary sampling units, is highly exceptional. The general conclusion is not that some analysis is generally to be recommended but that any investigator proposing to use ancova on Design 2 must reason carefully to settle upon the proper analysis.

Notes for Section 8

p. 8.8 [1] Since we calculated from the adjusted scores, the computer printout showed 370 d.f. , and a higher F . Taking the number of degrees of freedom as 1, 369, the F ratio is significant at the .01 level.

p. 8.14 [2] This could be called b_t , comparable to the β_t of Section 3.

9. Multivariate considerations

In the course of this project we made several multivariate analyses. and gained some experience in thinking about the decomposition of multivariate relations. I shall discuss those materials only selectively and briefly, since consensus regarding univariate analysis needs to develop before we try to resolve multivariate issues.

Simple correlations

Since Robinson and Yule and Kendall, it has been recognized that correlations change in going from the aggregate to the individual. I am interested in a comparison across and within groups, whereas previous writers contrasted correlations across groups with correlations across individuals regardless of group.

The kind of issue that arises for the psychologist can be seen if we consider convergent and divergent thinking. A good many investigators have argued about the degree to which these are correlated, and correlations ranging from zero to fairly large positive values have been reported. Those correlations have typically been calculated by measuring schoolchildren in a number of classes and pooling all cases. It is reasonable to suppose that the classroom can have an effect on the level of divergent thinking (D) for children who stand at the same point on convergent thinking (C); Torrance and others expect certain tactics of teachers to inhibit divergent thinking. If the teacher's effect on D is uniform over the range of D, and unrelated to the class mean on D, $\beta_{DC(w)}$ will exceed $\beta_{DC(b)}$. The overall regression coefficient will fall between them. The three correlations will similarly be discrepant. It would appear, then, that an attempt to sort out within- and between-groups relations is necessary to pursue any argument about the structure of abilities. However, the within and between relations differ because of demographic effects when group membership has no causal consequences. Computing separate correlations or regressions adds information but leaves interpretation equivocal.

Correlations of reading outcomes. In exploring the Cooperative Reading data our eyes were caught by the correlations between subtests of the Stanford Achievement Test within the LE and B treatments. In the conventional correlation matrix (over all individuals) the correlations of Spelling with other subtests were conspicuously lower in the B treatment. This could be of substantive interest. The LE program is, on its face, a more integrated approach to language and as such would perhaps generate higher correlations among outcomes than the Basal method.

¶ The obvious next step was to decompose the correlations. A typical set of values is that relating Spelling to Word Reading:

	Within LE	Within B
Conventional	0.76	0.54
Between-classes	0.90	0.83
Within-classes	0.61	0.39

This result, and others, seemed to indicate that the treatment chiefly affected within-classes correlations. The fact that between-classes correlations were consistently large is also of interest. Although correlations of aggregates often are large, it would be possible for teachers to vary the proportionate emphasis they give to different outcomes, and if so the between-groups correlations would fall off.

Correlations are affected by variances, and if the groups were selected differently in the two treatments, or moved farther apart in one treatment than another, this could account for differences in correlations. In fact, the within-classes variances proved to be much the same across treatments in all subtests except Spelling. In Spelling, the within-classes variance for B was more than double that for LE. This may be an important substantive finding, and one that is less striking in the conventional analysis. Here is the full set of variances:

	Reading		Spelling	
	Within LE	Within B	Within LE	Within B
Conventional	47.65	47.67	32.62	48.87
Between classes	22.28	16.32	17.81	15.84
Within classes	25.40	31.36	14.82	33.06
Estimate of η^2	0.46	0.34	0.55	0.32

The higher intraclass correlations for LE are consistent with the somewhat higher intraclass correlations for LE on the Pintner pretest (0.37 vs. 0.23).

We regressed Reading on Spelling and Spelling on Reading, obtaining these coefficients:

	Reading on Spelling		Spelling on Reading	
	Within LE	Within B	Within LE	Within B
Between classes	1.06	0.84	0.81	0.82
Within classes	0.84	0.38	0.47	0.40

The one clear finding is that between-classes regression slopes are considerably larger than within-classes slopes. Similar discrepancies were found for other pairs of variables. This finding is not, I think, to be dismissed as a consequence of greater measurement error in the individual scores. Rather, it is a statement that between-class differences in measured achievement are highly stable across outcomes in these elementary schools. Perhaps a part of the higher relation arises from conditions of test administration; correlated errors due to high or low group morale would make the regression steeper. More likely, the crucial fact is that individual patterns of difficulty -- the good reader who is weak in spelling and his opposite -- lower the within-class correlation but balance out over the class. Between-class differences in one subject would not be predictable from differences in another if there were a strong tendency for one teacher to put more emphasis on Reading (relative to Spelling) than the next teacher, or to have greater success in teaching one subject than another.

One might, as he prefers, emphasize the similarity of the Spelling-on-Reading regressions across treatments or the dissimilarity of the Reading-on-Spelling regressions. The proper conclusion appears to be that (at least in the samples) the within-groups joint distribution in B is distinctly different from that in LE, the former having a much greater dispersion. Another way of summarizing the same information is to emphasize the difference in η^2 (greater in LE for both variables). Since (perhaps fortuitously) η^2 on the Pintner pretest was greater for LE, interpretation must be left open.

The methodological moral of this exercise is that correlations among variables may be calculated within and between groups, but should not be interpreted by themselves. The information of importance is contained in the joint distribution of X,Y means and of X,Y deviations expressed in a uniform X metric and a uniform Y metric. Assuming that all distributions are normal, three parameters describe each distribution shape (and two parameters describe location). Correlations are derived from standardized measures, and the standardization of a variable is different for each distribution; contrasts are invariably distorted.

For similar reasons, use of unstandardized path coefficients generally is recommended. Because a direction of relationships has been postulated, interpretation is simpler than in the Reading-versus-Spelling example. Path coefficients have often been calculated from disaggregated data. It appears advisable to partition the structural regressions, making between-groups and within-groups analyses, despite the probable equivocality of the findings. (See also p. 9.23.)

The comments made here apply to Härnqvist's analysis of relations outcomes in the International study, mentioned on p. 2.12. He not only shows some striking differences among correlations at the individual and aggregate levels but makes the suggestion that the disaggregated correlations be recomputed for individuals within schools and schools within countries. He sees the lines he has opened up as dealing with some highly significant substantive questions. If my reasoning above is correct, the correlations ought to be supplemented by the pertinent variances, to give a sense of the joint distributions. Only this can give the reader a basis for interpretation.

Component analysis and factor analysis

Variables are reorganized into components or factors for three purposes: (1) Orthogonalization. Even if all the information from the original variables is retained in the orthogonal variables, it is often easier to carry out calculations and to make plots and summary statements in terms of orthogonal components. The simplest case of this kind of simplification is the change of variables X and Y to the set X and Y·X, the latter being a partial variate. (2) Rank reduction. There is redundancy in almost any set of variables, and the set can perhaps be compressed to fewer variables without much loss of information. Use of a limited number of components or factors simplifies, and relationships involving fewer variables will often crossvalidate in new samples better than relations fitted to a large number of variables. This is why minor factors or components are ordinarily discarded. (3) Identification of constructs. The purpose of rank reduction is to arrive at simple, stable empirical statements. The purpose of rotation of the factor set is to arrive at simple, stable descriptive or theoretical propositions. Those who rotate a set of such variables are searching for what are often called "underlying dimensions". Perhaps it is better to think of these as constructs, as working hypotheses regarding variables that can be used to formulate a satisfying theoretical network. If a good set of variables is found, relationships can be summarized in sentences that are comparatively simple, in the sense that each proposition employs only a few constructs in the set -- even though all the constructs are important enough to enter some sentences. (The reader may recognize Thurstone's concept of simple structure as an illustration of this desideratum.)

The partial variate Y·X is defined as $Y - \beta_{YX}X$. Since we have seen that the regression coefficients from between, within, and overall analyses differ, three distinct partial variates will be formed by them. The Y·X formed in an overall analysis will not ordinarily be uncorrelated with X either within groups or between groups. More generally, a variable set that is mutually orthogonal in one of the three analyses will almost certainly not be orthogonal in the other two.[1] Insofar as an investigator is primarily interested in orthogonalization, then, he may need separate orthogonalizations for the between- and within-groups segments of the data. As a minimum, he must decide which set of intercorrelations he wishes to reduce to zero.

A similar comment is to be made regarding rank reduction. When the first n_f dimensions from a larger set of n_v variables are retained and the remaining information discarded, this process will discard a fraction of the between-group information and a fraction of the within-group information. Those two fractions may differ in amount and in character. Suppose that the analysis is made within groups, and the first three factors retained. Those factors may account for 80 per cent of the total variance within groups on all variables; they may account for 92 percent of the total variance between groups, or 70 per cent. When an overall analysis is carried out, the first component may arise largely from between-groups variance, or largely from within-groups variance, or from a mixture of the two in any proportion. The same is true of each later component. It is unlikely that several successive factors would arise from the same single source, unless the groups were formed at random and the between-groups information is nothing but noise. The practical implication is that a person who reduces his data on the basis of a single factor analysis at any one of the three levels retains the major fraction of the information at that level, while perhaps discarding a significant fraction of the information at one of the other levels. Where rank reduction is the aim, it is important to examine separately the within-group and between-group residual covariances or correlations, to make certain that they are negligible.

The most intriguing problems arise in the attempt to establish constructs on the basis of factor analysis. As was said earlier, variables that have the same operational definition may have different substantive interpretations at the individual and the aggregate levels. A factor is a weighted composite of observables (or variables that are in principle observable), hence the preceding statement applies to any factor. A composite that enters into simple between-groups relationships may have quite different relationships with corresponding individual-level variables (within groups or overall). The person using factor analysis as a tool in theory construction, then, will need one set of factors for his between-groups theory and another set of factors for his within-groups theory. To be sure, he may find that the two sets of constructs coincide, but that is a possibility to be evaluated, not assumed.

Discriminant analysis is the one context where separate multivariate decompositions have often been made within groups and between groups. Discriminant analysis is an attempt to describe differences between groups in terms of one or a few variables. In Fisher's famous example of two species of iris, a number of physical measurements were made on many specimens of each species. (More than two species could have been investigated.) The analysis reduced the measures to two composites which were sufficient to classify plants into the two populations with few errors. The first discriminant function is whatever composite has the largest intraclass correlation. The second is whatever composite of the remaining information has the largest intraclass correlation. And so on. As a first step in the analytic procedure, the within-groups variance-covariance matrix is factored into orthogonal components and these are standardized (within groups). A consequence of this standardization is that when any dimension is partialled out (as in going from the first discriminant function to calculation of the second), the multivariate distribution of residuals can again be described by orthogonal variables with unit variance. The means of the groups on the components are formed, and the first principal component of the between-groups covariances becomes the first discriminant function. Successive principal components become successive discriminant functions. The first two or more discriminant functions can be rotated if that is thought to give a more "meaningful" description of group differences. In the study of irises or other similar pools, the rotated discriminants might suggest something about the character of the genotypic differences between species.

A rather large number of factor analyses of aggregate data have been made. R. B. Cattell (1949, 1952) suggested that a group has a "syntality" analogous to the personality of an individual, and he paralleled his studies of dimensions of individual differences with some factor analyses of group differences. For other summaries or discussions of aggregate-level factor analyses, see Janson, 1969; Cartwright, 1969; and Tryon and Bailey, 1970. So far as I know, only Slatin (1974) has carried out factorizations of the same data at two levels. I discuss his study below.

Whether an investigator should want factors for between-groups and within-groups variance is a subtle decision; in some contexts, factors from an overall analysis are no doubt appropriate. When groups were formed by aggregation rules that are irrelevant to the matter at issue, the between-groups factors that reflect those rules will be of no importance. In Fisher's study, on the other hand, the aggregation represented a judgment that two pools of plants were distinct biologically. Therefore the pooled set of individuals represented an arbitrary mingling of between- and within-groups information. The causes of variation between groups were probably not the causes within groups, and the structure might well have differed from one group to another. The psychologist has most often regarded effects as strictly individual. Even Cattell, in his studies of individual traits, has analyzed overall correlation matrices, ignoring groups. This may be appropriate in some circumstances but probably not in all. If it is true that teachers affect scores on divergent thinking or spelling, to mix class-level differences into a study of individual differences gives us indirect and clouded information about individual growth in ability patterns. On the other hand, to use within-groups information as the basis for similar conclusions is a dubious practice, insofar as arbitrary or irrelevant assignment rules restricted the range on some variables and modified the intercorrelations.

We move now to an illustrative factorization of ability tests which will give some concreteness to what has been said. This set of data was examined some time ago, as an exploration. It was a poor choice from a substantive point of view. The tests were given to first-graders early in their school careers, and the between-group differences reflect neighborhood differences or rules for assigning children to classes rather than psychological causes. The between-group information is essentially a summary of demographic effects. Despite the likelihood that the overall analysis probably answers the questions most likely to be asked about these data, much can be learned from the contrasts among the analyses.

Analysis of correlations. Miss Webb analyzed three correlation matrices (overall, between classes, within classes) for eight pretests in the Bond-Dykstra data, using a total of 1049 cases from the B and LE treatments combined. (We had no reason to consider treatments separately.) Analyses were made with unity in the diagonal and also with estimated communalities; no insight will be lost by discussing just the latter at this point.

The first fact of interest is the high degree of multicollinearity in the between-groups correlation matrix, so high that the communality for the

Pintner score was 0.99. (As a consequence, the computer's attempt at varimax rotation produced a nonsensical result.) I rotated factors II and III in the within-groups analysis through 45°, to bring them more nearly into line with the corresponding factors of the other two analyses.

Table 9.1 presents the factor loadings, communalities, and percentages of variance accounted for by the first three factors in each analysis.

[It will be noted that the communalities were considerably higher between groups than [within groups, except for bCopying and bIdentical Forms which had large unique factors. Correspondingly, the common factors accounted for a larger fraction of the between-groups variance than of the within-groups variance.

The reader may plot the loadings for himself. He will see that the structure in the conventional analysis corresponds far more closely to the between-classes analysis than to the within-classes analysis. This is true even though the intraclass correlations were only about 0.30 (see below). Factors I and II in the conventional and between-groups analyses plot out as a quasi-simplex. The order in which the tests string out is identical except for the position of Listening. The within-classes analysis produces a two-cluster configuration: wCopying, wIdentical Forms, and wPintner fall along one vector and the other five tests cluster on another.

The table codes the tests differently in the three factor analyses to remind us that the between-classes and within-classes analyses look at

Table 9.1. **Three factor-analyses of readiness measures**

Conventional					Between classes					Within classes				
	I	II	III	h^2		bI	bII	bIII	h^2		wI	wII	wIII	h^2
Pintner	.7	-.2	-.2	.68	bPintner	.8	-.3	-.4	.99	wPintner	.8	-.4	-.1	.75
Phonemes	.7	.0	.2	.55	bPhomemes	.8	-.1	.1	.68	wPhonemes	.6	.2	.2	.48
Letter Naming	.7	.2	.2	.54	bLetter Naming	.8	.2	.2	.70	wLetter Naming	.6	.2	.3	.49
Learning	.6	.4	.1	.49	bLearning	.7	.5	.1	.74	wLearning	.5	.2	.2	.32
Copying	.4	.2	-.2	.28	bCopying	.3	.5	.1	.35	wCopying	.5	-.2	.1	.32
Identical Forms	.4	.1	-.2	.26	bIdentical F.	.4	.2	-.1	.18	wIdentical F.	.5	-.2	.0	.25
Word Meaning	.6	-.4	.2	.52	bWord Meaning	.7	-.6	.3	.90	wWord Meaning	.5	.2	-.4	.49
Listening	.5	-.3	-.1	.47	bListening	.7	.0	-.1	.55	wListening	.5	.1	-.3	.29

different variables. bPintner is the class mean (standardized after averaging), and wPintner is the individual deviation from that mean (likewise standardized). Because of standardization, the Pintner variable of the conventional analysis equals η^2 bPintner + $(1 - \eta^2)$wPintner, where η^2 is the intraclass correlation for Pintner.

Factoring covariances. The standardizing operation will distort information. If a variable has a small intraclass correlation, it has a small between-classes variance yet it has as much "weight" in a between-classes factor analysis of correlations as a variable with large variance. Conversely, a variable with little within-classes variance is given heavier weight in the within-classes analysis when correlations are used. Our next step, then, was to partition the overall correlation matrix of the scores into between-groups and within-groups covariance matrices, and then to factor those matrices.[2] We started with correlations because there seemed to be no reason for weighting one variable differently from another in the overall analysis; one could, however, start with the covariance of raw scores or of scores rescaled in some preferred manner. Any such scaling decision affects which variables dominate the first principal components of the overall analysis. Having begun with ones in the diagonal of the overall matrix, we had intraclass correlations in the diagonal of the between-groups matrix and values of $1 - \eta^2$ in the within-groups diagonal.

In the between-classes analysis, three factors accounted for 79 per cent of the variance, and little variance remained to be accounted for in subsequent factors. Therefore, Table 9.2 presents only the first three factors.

Table 9.2. Analysis of covariance matrices for readiness measures

Between classes						Within classes							
	bI	bII	bIII	h^2	η^2		wI	wII	wIII	wIV	wV	h^2	$1-\eta^2$
bPintner	.22	-.18	.02	.22	.30	wPintner	.63	-.07	.31	-.02	-.15	.52	.70
bPhonemes	.25	-.15	-.03	.25	.32	wPhonemes	.58	.24	-.18	-.14	-.26	.51	.68
bLetter Naming	.23	-.03	-.04	.21	.29	wLetter Naming	.57	.30	-.17	-.05	-.09	.46	.71
bLearning	.22	.11	-.01	.20	.31	wLearning	.47	.38	-.26	.26	.33	.61	.69
bCopying	.25	.58	-.19	.50	.54	wCopying	.35	.04	.24	.06	-.26	.25	.46
bIdentical Form	.15	.07	.53	.34	.35	wIdenticalForms	.43	.05	.52	.09	.30	.56	.65
bWord Meaning	.19	-.25	-.14	.23	.31	wWordMeaning	.50	-.30	-.13	-.50	.26	.67	.69
bListening	.13	-.03	-.01	.09	.16	wListening	.54	-.60	-.23	.35	-.04	.83	.84
% of variance	46	21	12	79		% of variance	39	14	11	9	8	81	

at 9.13

In the within-groups analysis, variance was extracted more slowly, and five factors are retained. Neither the b nor the w analysis was particularly close to the overall analysis of covariances (not shown here); it could be described roughly as "halfway between" the two.

The η^2 was particularly low for Listening, and particularly high for Copying. Possibly an explanation could be constructed from a search for ceiling and floor effects or other anomalies; alternatively, the patterning could reflect something about neighborhood characteristics. Since the children were tested near the start of the first grade, causal "class" effects are highly unlikely, except as irregular administration of tests affected class standings.

The three chief components of the between-groups covariance matrix are not much like the first three components of the correlation matrix. (Some of this shift comes about because we decomposed the entire covariance, rather than just the common-factor portion as before.) The general factor runs over all tests about equally, except for Listening and Identical Forms. Components bII and bIII are essentially specific to bCopying and bIdentical Forms.

The within-groups analysis shows fairly strong common factors. Listening loads more heavily than in the between-groups analysis, because of its small intraclass correlation. The first factor within groups is present about equally in all measures. Rotation could bring out the connection of wPintner with wCopying and the connection of wPhonemes with wLetter Naming, but the structure is not strongly patterned.

Slatin's analyses. Slatin (1974) factored 10 variables at the group and individual levels. His data were measures of ability and family background for boys, plus two indices of property value for their neighborhoods. His aggregates were areal units, such that the 516 boys were successively clustered into 47, 21, and 10 areas. Slatin factored each of four correlation matrices, extracting and rotating three factors. He was impressed by the differences between the factor structures, and suggests tentatively that a more sociological (more "social") explanation of phenomena will be reached when aggregate data are factored than when individual data are factored.

Our work perhaps sheds some light on Slatin's findings. The most striking change in going from the individual analysis to the aggregate analysis was a shift in loadings for Age. Pairing up the varimax-rotated factors of the four analyses, the loadings for Age are

	I	II
516 individuals	-.22	-.10
47 smallest areas	.03	-.74
21 medium areas	-.19	-.92
10 largest areas	-.24	-.95

Factor I is an ability factor and the chief markers for II, other than age, represent neighborhood wealth or father's status. Age had a much lower intraclass correlation than several other variables (Slatin, 1969), and hence, when variables were restandardized, tiny and fortuitous covariances across groups were inflated to the point of making Age a powerful influence in the aggregate-level analyses. (One covariance of about 0.01, I estimate, was inflated into a correlation of 0.72.) If covariances had been analyzed instead, the analyses at successive levels would have changed only gradually. At the individual level, there is a near-simplex running through the ability measures, then around to lot size and value of dwelling unit (DU). (Age and Delinquency have such low correlations that they do not enter the simplex.) Much the same simplex appeared in the aggregate covariance matrices.

If covariances had been factored, I believe that the only change would have been a reduction of the spread of the vectors; at the highest level of aggregation IQ and DU correlated 0.67.

It might have been wise for Slatin to have examined factors within aggregates. The relation of individual characteristics after neighborhood is held constant might be the best way to bring more purely psychological relations to the surface. But it might be equally interesting to decompose "upward", factoring values of $\overline{X} \cdot X$ to see what "social" factors might appear after the individual information was removed.

Suggestions.

Insofar as investigators seek only to replace original variables with a smaller number of composites that carry most of the differential information, no serious problems arise. There is no reason to try to establish homology between factors at various levels, and one will of course factor at whichever level he is interested in. His only major decision will be whether to standardize variables at that level or to use some other metric.

It is when factors are to be regarded as constructs that interpretation becomes awkward. It is unreasonable to expect homology. If only for statistical reasons, different results are likely to appear at the between-groups and within-groups levels. The grouping variables that modify the regression coefficients (as shown in Section 4) also modify the covariances, even when no causal effects are associated with the groups. Beyond this, however, original variables take on different meanings when aggregated. They can be expected to cluster differently, with the consequence that different constructs will be appropriate between and within groups. Let us consider total yields of wheat and potatoes (rather than per-acre yields). More agricultural counties will have larger yields of both potatoes and wheat. But within counties the farmer decides whether to plant potatoes or wheat, so the two yields may be negatively correlated at the level of the farm. In some problems it may make sense to track down just this patterning; in other problems where group boundaries seem to have little causal significance an overall analysis will suffice.

Brunswik's "ecological" ideas will help in interpretation. Any result obtained from a sample of persons is also representative of the sample of subecologies in which their behavior developed. If one samples groups and measures everyone in those groups, the correlational information is a statement about the distribution of behavior within and between groups, hence within and between subecologies. The results generalize over the population of groups sampled. When sampling is at the individual level, perhaps from a census roster, the result holds for persons who have grown up in a culture, distributed over its subecologies. The only difference from the study where groups were sampled is that with individual sampling there are too few persons from any one subecology to warrant examining its specific characteristics. From this point of view, then, the overall correlation describes an ecology in the large, the between-groups correlation contrasts subecologies within that ecology, and the within-groups correlation describes typical relations within subecologies. The factor analyst who intends to study a purely psychological question about the distribution and covariation of abilities is inevitably reporting on a phenomenon that is cultural, demographic, and sociological in part.

The overall covariance being a composite of between-groups and within-groups covariances, the adequacy of the overall data depends on the adequacy of the data at the two levels. Our Bond-Dykstra analysis used 1049 cases, and that is ordinarily enough to satisfy any factor analyst. But the information for the important between-groups portion of the covariance comes from a sample of 57 classes. Few would consider 57 cases a sufficient sample for a factor analysis.

If we assume substantial homogeneity of relations within the several classes -- which is probably necessary if a within-classes factor analysis is to be taken seriously -- then the within-groups covariances are well established when 50-odd classes are pooled. Yet insofar as Table 9.1 is representative, the overall analysis that has conventionally been made rests far more on the fallible between-groups covariances. Now if within-class relations are not homogeneous, the whole analysis is suspect. The pooled-within-class analysis has uncertain meaning, and its stability is a function of the number of classes, not the number of individuals.

The case for considering separately the between and within factor analyses is especially impressive when we move out of the ability domain. An example is the Learning Environment Inventory (LEI) developed by G. J. Anderson and Walberg (1974) within Harvard Project Physics. This is a collection of items describing instructional procedure and student attitudes; the student responds so as to report how he perceives his class. Items have been assembled into scales on the basis of their intercorrelations, and the scales have been factor-analyzed. Insofar as correlations arise simply from semantic overlap of items, one would expect similar joint-distribution shapes within and between classes. But the correlational structure, insofar as it reflects psychological differences, may be quite different. Within the class, the correlation reflects the correspondence of perceptions. Are the students most prone to describe the class as "apathetic" also the ones prone to describe it as "difficult"? Across classes the correlation speaks of a different phenomenon: When the students collectively describe the class as "difficult", do they also describe it as "apathetic"? The former refers to the phenomenology of the student, compared to other students rating the same events. The latter refers to behavioral differences between classes (though some of those differences are perceptual rather than objective).

The purpose of the LEI is to identify differences among classrooms. For it, then, studies of scale homogeneity or scale intercorrelation should be carried out with the classroom group as unit of analysis. Studying individuals as perceivers within classrooms could be interesting, but is a problem quite separate from the measurement of environments.

Empirical test construction. Once the question of units is raised, all empirical test construction and item-analysis procedures need to be reconsidered. Is it better to retain items that correlate across classes? or items that correlate within classes? A correlation based on deviation scores within classes indicates whether students who comprehended one point better than most students also comprehended the second point better than most -- instruction being held constant. A correlation between classes indicates whether a class that learned one thing learned another, but this depends first and foremost on what teachers assigned and emphasized. It is the items that teachers give different weight to that have the greatest variance across classes. If some teachers work hard to teach use of the semicolon and others consider it unimportant, the semicolon items will correlate comparatively high over classes. If teachers who care about semicolons may or may not care about colons, a low across-classes correlation between semicolon and colon items is to be expected. This leads us to regard the between-group and within-group correlations of items as conveying different information, and makes the overall correlation for classes pooled an uninterpretable blend.

Multiple regression and related techniques

A school-effects model. It makes sense to consider two or more measures on the individual or the collective for many purposes. A simple school-effects study might include X_1 = family background, X_2 = ability of teachers, and Y = student achievement. Suppose that data in just one community will be examined. It appears best to identify the X_2 of the student's own teacher. A conventional analysis might evaluate

$$\hat{Y}_p = \beta_{1\cdot 2}X_{1p} + \beta_{2\cdot 1}X_{2p} \tag{9.1}$$

A more sophisticated individual-level analysis might add contextual variables as a last step:

$$\hat{Y}_p = \beta_{1\cdot 2}X_{1p} + \beta_{2\cdot 1}X_{2p} + \beta_{3\cdot 124}X_{1s} + \beta_{4\cdot 123}X_{2s} \tag{9.2}$$

Here, X_{1s} and X_{2s} are school-level aggregates.

The fact that the predictors -- especially X_{1s} and X_{2s} -- tend to be correlated creates difficulties of interpretation. The difficulties are exacerbated by the fact that the number of schools is small. Consequently, $\rho_{1s,\ 2s}$ is likely to vary considerably from community to community. This is a fact about local affairs, acceptable enough in considering fixed schools in a fixed community. If communities are compared or an attempt is made to generalize over communities, it is highly likely that the β that one wishes to interpret will be much affected by the value of ρ. Particularly if $\rho_{1s,\ 2s}$ is large, the $\beta_{3\cdot124}$ and $\beta_{4\cdot123}$ can be compared only at considerable risk. It is not at all unlikely that their balance would shift in another year in the same community even if $\rho_{1s,\ 2s}$ does not change.

A decomposition seems to require aggregation at the class level (c) as well as the school level (s). In place of (9.2) we have three equations:

$$(9.3) \qquad Y_s = \beta_{s1\cdot2} X_{1s} + \beta_{s2\cdot1} X_{2s}$$

$$(9.4) \qquad Y_c - Y_s = \beta_{c1\cdot2}(X_{1c}-X_{1s}) + \beta_{c2\cdot1}(X_{2c}-X_{2s})$$

$$(9.5) \qquad Y_p - Y_c = \beta_{p1\cdot2}(X_{1p}-X_{1c}) + \beta_{p2\cdot1}(X_{2p}-X_{2c})$$

Assume that all definitions of parameters take number of students into account. What is here written as $\beta_{s1\cdot2}$ could be written $\beta_{Y1s\cdot2s}$; in the notation of (9.2) it would be $\beta_{3\cdot4}$ -- with no partialling of 1 and 2. The last term in (9.5) is entered pro forma. Teacher quality X_{2c} cannot be disaggregated, hence $X_{2p} = X_{2c}$ and the term vanishes. A global school quality would vanish from c and p equations.[3] The sums of squares from (9.4) and (9.5) can be combined into components of the within-schools SS.

The correlation ρ_{1s2s} is now relevant only to (9.3). The usual problem of allocating variance between two correlated predictors (p. 2.17) arises at the school and class-within-school levels, but ρ_{1s2s} applies to one and ρ_{1c2c} to the other. In general, of course, interpretation of an equation at the within-class level takes ρ_{1p2p} into account.

The important point to remember here is that $\beta_{s1\cdot2}$, $\beta_{c1\cdot2}$, and $\beta_{p1\cdot2}$ are coefficients for different predictors, predicting different outcomes.

The overall analysis of (9.1) evaluates only two out of five parameters; even the analysis with added contextual variables leaves the components entangled. From an explanatory point of view a single equation fitted to the ecology provides less information than the set of equations.

Partialling. A multiple-regression equation gives regression weights for one variable with others held constant. A partial correlation relates two residualized variables. What usually goes unrecognized is that the variable carrying the same label becomes an operationally different variable at each level of aggregation. If we form $Y_c \cdot X_c$ at the aggregate level for some group c, that value will rarely equal the average for c of the $Y \cdot X$ formed by partialling at the individual level.

Härnqvist (1975, p. 102) reports partial correlations (at the end of secondary school) for achievement in literature with achievement in science with reading comprehension held constant:

	Iran	England
Individual level	0.25	-0.13
School level	0.36	-0.60

This is certainly of interest, as a kind of documentation for the "two cultures" stereotype of British education. Where we must be cautious is in believing that the same pair of variables has been correlated in each instance. The variables may be denoted (with some notation that should be transparent) as $L - b_1 Rd$ and $Sc - b_2 Rd$. But both b_1 or b_2 take on a new definition and a new numerical value in each cell, as follows:

$$\begin{matrix} b_{tI} & b_{tE} \\ b_{bI} & b_{bE} \end{matrix}$$

It seems to me highly dangerous to compare variables across levels and across collectives when the operational definitions shift. To argue that the several operations represent the same construct seems to entail an enormous burden of proof; in one context reading may be a proxy for individual SES and in the next context it may be more a proxy for global school quality. This argument applies obviously to path analyses, since most of their predictors are partial variates.

I can make only one recommendation and it is no more than a palliative. Recall that in the Anderson reanalysis (p. 5.8) Webb and I defined a variable ABIL = ABILITY - 0.47 PRECOM; where 0.47 was the overall regression coefficient of ABILITY on PRECOM (within treatments pooled). We entered this variable in the within-collectives and between-collectives analyses. The correlation of ABIL with PRECOM in each of these analyses was close enough to zero that we were able to reach interpretations; we did not have a mind-boggling shift in definition. I speak of this procedure as a palliative because one is not guaranteed that in each subset of the data the variable so formed will have a low correlation with the variable whose contribution was adjusted downward.

Table 9.3. Zero-order and multiple regression coefficients for Head Start classes at two levels of aggregation, with Metropolitan Readiness Test as dependent variable

	POPED	POPINC	POPOCC	NKIDS
Between centers				
Zero-order	3.47	6.06	6.72	0.72
Multiple	3.21	4.21	2.40	2.18
Within centers				
Zero order	3.89	1.56	1.39	-0.34
Multiple	3.61	0.50	0.58	-0.01

Based on Smith & Bissell, 1970

The multiple-regression equations given by Smith and Bissell, two of which appear in Table 9.3, are further evidence on the same point. The between-centers equation tells a quite different story, on its surface, than the within-centers equation does. All variables contribute to the former equation; within centers, only POPED has a large weight. (All variables had been rescaled to similar metrics.) Some of the disparities are explainable in terms of the zero-order coefficients, but the shift in NKIDS is not. The difficulty lies in the fact that the weight is for "NKIDS with the other three variables partialled out". Since the correlations among predictors changed from one level to another, the partial variates are radically different in their definitions from one analysis to the other.

Disparities such as these (including disparities across treatments) caused no difficulty for Smith and Bissell, since they were not attempting a causal interpretation of regression weights. Sociologists often do attempt such interpretations, however. Some set of operationally defined quasiorthogonal composites would appear to be necessary for any comparison of regression coefficients across levels or across treatments. One might be wiser to examine zero-order regression coefficients for such composites than to use multiple regression, but path analysis calls for multiple regression.

Notes for Section 9

p. 9.7 [1] Duncan et al. (1969, p. 54) establish the following relationship:

$$r_t = (\eta_X^2 \eta_Y^2)^{1/2} r_b + [(1 - \eta_X^2)(1 - \eta_Y^2)]^{1/2} r_w$$

p. 9.13 [2] A note on procedures may be helpful to some readers. The SPSS programs "do not accept" covariance matrices, but since the analytic routines for correlation matrices apply to covariance matrices, we fooled the computer into thinking it was analyzing correlations. We entered the between-groups matrix with ones in the diagonal, used an option in the program to instruct the computer that the "estimated communalities" were the vector of η^2, and called for a principal-factor analysis with those communalities. The result was a principal-components solution for the between-groups covariance matrix. The same method was used for the within-groups matrix, with the vector of $1 - \eta^2$.

p. 9.21 [3] I noted earlier that product terms can be added, particularly to enter global variables in lower-level equations. Thus (9.5) could become

$$\hat{Y}_p - \hat{Y}_c = \beta_{p1}(X_{1p} - X_{1c}) + \beta_{p1c2}(X_{1p} - X_{1c})X_{2c} \tag{9.5a}$$

These two predictors are uncorrelated. A positive β_{p1c2} implies that the abler teacher is associated with a steeper Y-on-X regression within the class. Even though analysis is based on individuals it is the number of teachers that regulates the sampling error of this coefficient.

10. The Road Ahead

This report has been chiefly concerned with educational research in which data are collected on students with classrooms or on classes within districts. The difficulties noted in a great variety of commonplace studies (Table 10.1) imply a need for new strategies of design and interpretation. It is not clear to me just which of those difficulties will trouble investigators in other fields -- for example, students of voting behavior or of reference-group theory; my impression is that their difficulties include those treated here plus additional ones.

I started with a concern for a somewhat specialized kind of research, the study of Aptitude × Treatment interaction. That kind of inquiry has in the past examined overall regressions of outcome on aptitude for pools of students assembled from many classes. According to the analysis here, no meaningful question is being asked about interactions in group instruction unless between-group and within-group regressions are considered separately. The implications of my explorations extend far beyond the studies of interactions, however.

Every educational or sociological study that attends directly to regression coefficients and correlations, including studies using structural-equation models, must be thought through in the light of the argument I have presented. Sometimes a traditional analysis looking only at overall relationships or between-group relationships will prove to be adequate for the purposes of the investigation. More often, I suspect, the investigator will find that a more complex decomposition will add to his store of information. But it will also make him painfully sensitive to the vagaries of results -- no matter how analyzed -- when only a limited number of groups are sampled. And it begins to appear that even the best of analyses will leave the causal interpretation

Table 10.1. Kinds of investigation within education, psychology and sociology where difficulties have been identified

	Page reference in this report
Analysis of covariance	1.3a ff., 1.17a, 8.1ff.
Aptitude × Treatment interaction	2.1 ff., 3.21, 5.1 ff.
Attenuation, correction for	1.7, 6.1 ff.
Classroom climate, assessing	9.18
Context effects	1.16 ff., 1.22 ff., 3.16
Correlation, simple	2.12, 9.1
Evaluating treatments	1.3a ff., 1.13, 1.15, 1.18, 2.17, 8.3 , 8.8, 8.11 ff.
Factor analysis and component analysis	9.6 ff.
Item analysis in test construction	9.19
Multiple regression	1.7, 9.20
Partial correlation	9.22
Path analysis	2.19, 3.17, 9.24
Placement rules, development of	2.8
Predicting scores within aggregates	1.12, 4.9
Regression, comparing across levels	1.17a, 3.16, *et passim*.
Reliability studies	6.1 ff.
School-effects studies	1.3a, 2.17 ff., 9.20
Social-area analysis	1.23

of results equivocal, unless assignment to groups was under the control of the investigator.

For the most part, I have made the following assumptions:

1. Every member participates in just one collective at the next higher level.

2. At successive levels of aggregation collectives are completely nested.

3. Aggregates at the highest level are random and membership at lower levels is fixed.

4. A member has no direct effect on scores of a collective to which he does not belong.

5. Data are complete at the lowest level of disaggregation.

6. There is a known causal order, X preceding Y.

The chief difficulties identified in educational studies that fit these assumptions are as follows:

1. There is no warrant for direct generalization from groups formed in one way to groups assembled in some other manner.

This requires revision of previous thinking about the classification of students on the basis of aptitude (Cronbach & Gleser, 1957; Cronbach & Snow, 1976). An investigator who applies a treatment to a number of individuals separately will identify a certain regression of outcome on aptitude, and may devise a selection rule so that only promising individuals will receive the treatment in the future. If there are two treatments, he may devise a classification rule for deciding which treatment an individual is to receive. These rules are a suitable basis for future decisions if individuals from the same population are to be treated individually. If they are to be assembled into instructional groups, however, the overall regression of outcome on aptitude need not be the same as before; moreover, the possibility of distinct within- and between-group relations should be taken into account in the decision rule. The policy based on the individual-level investigation provides only a tentative hypothesis about group-level instruction. The same argument applies if the original study uses group instruction. If the initial groups are assembled by a certain rule, the findings apply to future groups formed in the same manner. If future groups are formed in a different way, the old conclusions do not apply directly. In the short run, it appears necessary to insist on fresh validation research when persons are taught (or carry out tasks) in groups and the basis for forming groups is modified. In the long run, studies of groups formed in different ways might develop a theory that would permit reasonable predictions about kinds of groups that have not been directly investigated.

2. Most experimental studies carried out in classrooms have been analyzed by means of "individual level" (overall) statistics, with classes ignored. The between-groups regression of outcome on aptitude is likely to differ from that within groups; the overall analysis combines the two kinds of relationship into a composite that is rarely of substantive interest.

Individual-level analysis may be undertaken as a deliberate choice. Analysis of pooled individual data is warranted when:

a) The investigator (as in much survey research) is interested in a composite description of individuals who are mixed into groups in the population as they are in the sample, and not in a causal interpretation of group-related effects; or

b) The investigator is prepared to assume that any causal effects associated with group membership are trivial in magnitude; or,

c) Known conditions of group formation make the overall statistic a comparatively efficient estimator of a between-groups parameter.

3. A conscious decision, rooted in the theoretical background or the practical context of the investigation, is required to identify the appropriate units for sampling, assignment to treatment, and analysis. This is true in factor analysis, item analysis and empirical keying of tests, prediction of scores for lower-level units, etc. An individual-level analysis, a between-groups analysis, and a within-groups analysis address substantively different questions and usually give different results.

4. When analysis of covariance is used to compare instructional treatments, and data have been collected by using an intact collective (e.g., the school) as the primary sampling unit and the unit of assignment, the theoretically appropriate adjustment appears to be that given by the regression coefficient across such collectives (perhaps within blocks). In non-random experiments, analysis at some other level may give a considerably different estimate of the adjusted treatment effect.

To determine a between-classes or a between-schools regression coefficient with reasonable precision one must have a much larger sample of such units than is normally available to the experimenter. Consequently, his adjustment may be heavily influenced by sampling errors.

5. In the study of aptitude-treatment interactions, particular interest attached to the between-classes regression coefficient, because instructional treatments will most often be assigned to whole classes. As has been said, it is rarely practical to determine such a regression coefficient with precision.

This fact, together with the hazard in generalizing to groups formed by new assembly rules, seems to discourage altogether a sheer empirical search for ATI in classroom instruction.

6. For the within-groups coefficient, the sampling error depends on the number of classes and the variability of the specific within-groups coefficients. The customary manner of examining such sampling errors ignores that variability. Sometimes the estimate of the pooled within-groups coefficient will be undependable even though hundreds of individuals provide data.

7. No secure causal interpretation can be given to the between-groups regression coefficient, the within-groups coefficient, or their difference. The simple interpretations regarding "context effects" and "school effects" that have been made in the past are not defensible.

At least three causal processes may affect regression slopes: direct effects of the individual's characteristics on his performance, competitive effects or other differentiation of experience arising from the heterogeneity of individuals within groups, and processes that raise or lower the outcomes for groups with high (low) standing on the predictor variable. The first two of these are confounded in the between-groups coefficient, and the latter two affect the within-groups coefficient. There is no direct way, then, to evaluate the magnitudes of the three effects from the two observed coefficients (especially as the two components affecting a regression coefficient may work in opposite directions).

8. Even when only strictly individual causes operate, the between-groups and within-groups regression coefficients will generally differ from that calculated for individuals pooled, because of demographic differences among the groups. These differences may reflect social processes, or aggregation rules of the data collector. This creates great ambiguities in interpreting differences between parameters at two levels of aggregation, or differences between parameters for two treatment populations that were aggregated into groups by different rules.

9. Studies of regression coefficients, and analyses of covariance in which regression coefficients play a part, cannot be soundly interpreted without due consideration of errors of observation of the predictor variable. Direct comparison of an observed between-groups regression coefficient with an observed within-groups coefficient is surely unsound. Observed regression coefficients will not be patterned like the coefficients for universe scores; yet the latter are of more fundamental interest. Contrary to the usual belief, group-level information will not necessarily have a smaller standard error and a smaller coefficient of generalizability than information on individuals within groups.

The information that would be required to evaluate the reliability (generalizability) of group-level data has not been collected in studies to date. Indeed, the theory for such studies has barely begun to evolve. Yet -- let me repeat -- without proper disattenuation one arrives at incorrect answers to the kinds of questions educational research workers and sociologists have been trying to study.

Sweeping recommendations cannot be made because of our present ignorance and because tactics must be suited to each substantive problem. The following suggestions are derived in part from what others have written.

1. There should be a deliberate attempt to conceptualize the processes operating when persons are treated in groups or hierarchical structures formed on a certain basis. Once a process is postulated one can hope to suggest indicators of intermediate events that will give information on the strength of the process. Such more complete specification of the modelwill be required to get data that warrant causal interpretation.

2. The process by which individuals are assigned to (or voluntarily enter) groups should be specified as fully as possible. In general, when group membership is under full control of an investigator -- whether he uses a random process or groups persons having specified characteristics -- interpretation of findings will be freed of many of the uncertainties that arise with self-selected groups.

Where the collectives already exist, there should be a careful description of the way collectives differ. These discriminating variables are a part of the causal process by which group effects are generated, and they may create noncausal relationships at the group level.

3. Some amount of direct experimentation on context effects should be carried out, to supplement and lend supporting insight to the correlational studies under field conditions. It should be instructive, for example, to investigate how learning differs, cognitively and affectively, when persons work alone and when similar persons work in a group composed in one or another manner. To experimentally disentangle effects of group composition, of group differences in the treatment delivered, of within-group differences in experiences, and of individual differences properly speaking would inform both methodologists and theorists. Such studies will necessarily be limited in time and size, and cannot answer questions about cumulative effects of group experience.

4. In an experiment or survey, it would often be advisable at the first stage of sampling to select

units at the highest level at which causal processes of interest operate. In an experiment, it is those units that should be assigned to treatment. Thus, in a study of school desegregation, it seems reasonable to take the community as the sampling unit. This is true not only because desegregation plans are typically district-wide but also because the attitudes of patrons in one school are likely to influence their fellow townspeople. An exception to this largest-unit principle is noted for static descriptive studies such as public-opinion polls; if the interest is only in the population average or some similar statistic, conventional sampling of small units is efficient. Another qualification: In contrasting treatment effects at one level it may be efficient to select random collectives at the next higher level, and then to divide the members of each one among treatments.

Once the large unit is chosen, it may be sensible to sample smaller units within it: schools within a district, classrooms within a school, or students within a classroom. Such multistage sampling has to be designed to fit the purposes of the particular investigation (Jaeger, 1970). For the kinds of studies discussed in the report, it will almost always be more important to increase the number of schools or classrooms. It may be appropriate to test only a fraction of the individual students, if that economy permits an increase in the number of groups.

In obtaining a between-collectives regression coefficient, there are advantages in an extreme-groups design, most or all of the data being taken from groups whose means are far out on the predictor scale. This design has superior cost-effectiveness for evaluating a univariate regression equation. I do not suggest, however, that groups be formed by assembling individuals with extreme scores; such groups are not appropriate representatives of the reference population of collectives formed in a more normal manner.

5. In most studies, it will be impractical to collect extensive data on a large sample of high-level units. Not many investigators will be in a position to investigate 100 school systems, or even 100 classes. Research on higher-level units, then, will have to be more in the character of case

studies, and less in the character of statistical studies. This point was made by Merton and Kitt (1950) in one of the first modern papers on group effects, but it seems to have dropped from sight during subsequent attempts to draw conclusions about context effects from survey-like data.

6. The mode of analysis of effects is to be determined by the substantive model suggested for the processes at the several levels. It will often be appropriate to form separate structural models for between-groups relations and for within-groups relations, taking into account the rules by which collectives are formed.

7. Plots of the data should be made, to the greatest degree possible. Repeatedly, bivariate plots of group means have improved my interpretation of between-group statistics (often by inducing caution). Studies of groups will usually be limited in size, and outliers can make a large difference in the statistics.

8. In predicting scores of members of collectives, it will generally be sound to estimate the dependent variable for the collective and then add the predicted deviation of the member's score from that group value, instead of making a one-stage prediction.

Even if the two-stage prediction accounts for rather little additional variance, it may make statements about atypical individuals that differ appreciably from the simple prediction.

I have adopted the working hypothesis that treating persons in groups does make a difference. The makeup of a group may determine the events that impinge on the group as a whole, and may condition the events that impinge on the individual group member or on his perception of them. Some investigators will prefer a simpler working hypothesis that tries to explain the world without reference to group effects. No doubt group effects are negligible in some instruction and in some social processes, but even the investigator who prefers to deny their existence will be wise to allow his data to speak on the point, to the extent that a design of modest size can give information. It is intellectually legitimate to adopt a strong model that assumes absence of group effects, but this is likely to be a poor strategy in any substantive field where one has insufficient experience to make the assumption persuasive.

The issues that have come to light in this paper lead me to think that educational studies conducted in classrooms or with data from schools and school districts have almost never been analyzed correctly. Those few investigators who have taken the collective as the unit of analysis have rarely brought to the surface the potentially interesting within-group information. Moreover, survival of the null hypothesis with groups as the unit of analysis must often have been given a substantive interpretation without realization that the sample was insufficient to make the study informative. Per contra, use of the individual as unit of analysis when data are collective is likely to reject null hypotheses falsely. Descriptive as well as inferential statistics obtained by analyzing data on collectives at the individual level are open to misinterpretation, except where the interpreter realizes that he is looking at a composite figure for an assemblage of groups.

Oddly, the history of the aggregation problem in sociology, politics, and psychology has been one of regarding individual-level relationships as the information of primary interest, and group-level relationships as a distorted shadow of the former. Once the conventional individual-level information is seen as a composite of within-groups and between-groups effects (at least for certain variables and certain collectives), the situation is nearly reversed. The conventional mixture of the two effects is not usually the most meaningful variable to enter into hypotheses.

Not all studies in collectives will move in the direction emphasized in this report. There probably are kinds of investigation (factor analysis of reading-readiness tests being one) where the question can best be posed at the individual level even though data come from collectives. Even in such studies, however, the investigator would do well to ponder the proposition that if he randomly samples one student from each school in a large area he will get a different result than if he includes all the students in all the schools, or confines his analysis to students in a single school. Methodology should be matched to the substantive context; for example, factor analysis of the Learning Environment Inventory would seem particularly to call for distinguishing relations between and within groups.

The methodological maxim appears to be that an investigator who collects data on collectives ought to take an explicit position on the role of between-group, within-group, and individual-without-regard-to-group effects in the variables he studies. He may opt deliberately for any one of several analyses, but he should not back into one of the analyses merely because it is commonplace or convenient. Perhaps an investigator will wish to leave the question open, and analyze the data at several levels. This is to be encouraged so long as each analysis is logical, and the interpretation is ecological. To regard analyses at two levels as alternative ways to answer the same question is rarely if ever justified.

I have suggested elsewhere (Cronbach, 1975) that the ideal of establishing lawlike, lasting relationships in the social sciences may be unapproachable. In that paper I was focussing on the study of the psychology of individuals. This report makes the difficulties seem even more forbidding. A social science must deal with collectives, and the cost of obtaining data on collectives is great. It appears that the only recourse is to make more use of the data we can afford to collect, appreciating hints in the data with due regard for their uncertainty, and enriching our quantitative summaries with awareness of the qualitative context of the events.

References and Author Index

Alwin, D. F. Assessing school effects: Some identities. Paper presented to Pacific Sociological Association, 1975. (Revised draft used here.) 1.3, 3.19

Anderson, G. J., & Walberg, H. J. Learning environments. In H. J. Walberg (Ed.) *Evaluating educational performance.* Berkeley, Ca.: McCutchan, 1974. Pp. 81-98. 9.18

Anderson, G. L. A comparison of the outcomes of instruction under two theories of learning. Unpublished doctoral dissertation, University of Minnesota, 1941. 5.1 ff.

Barton, A. H. Bringing society back in: Survey research and macromethodology. *American Behavioral Scientist*, 1968, 1, 1-9. 1.17

Barton, A. H. Comments on Hauser's "context and consex". *American Journal of Sociology*, 1970, 76, 514-517. 1.17, 1.17b

Bell, D. The basic structure of knowledge: What we may no longer take for granted. Unpublished memorandum for a conference on philosophy of science at Aspen-Berlin, September, 1975. 1.20

Blalock, H. M., Jr. *Causal inferences in nonexperimental research.* Chapel Hill: University of North Carolina Press. 1964. 1.27

Blau, P. M. Formal organization: Dimensions of analysis. *American Journal of Sociology*, 1957, 63, 58-69. 1.23

Bock, R. D., & Wiley, D. E. Quasi-experimentation in educational settings: Comment. *School Review*, 1967, 75, 353-366. 2.17

Bond, G. L., & Dykstra, R. The Cooperative Research Program in first-grade reading instruction. *Reading Research Quarterly*, 1967, 2, 5-142. 5.11 ff., 8.8, 9.2, 9.11 ff.

Bowers, W. Normative constraints on deviant behavior in the college context. *Sociometry*, 1968, 63, 58-69. 1.17 ff., 4.10, 6.10

Bronfenbrenner, U. Experiment in human ecology: A reorientation to theory and research on socialization. Unpublished address, American Psychological Association, 1974. 2.16

Bronfenbrenner, U. The ecology of human development in retrospect and prospect. Paper presented to a Conference of the International Society for the Study of Behavioral Development, Guildford, England, July, 1975. 2.16

Bronfenbrenner, U. The experimental ecology of education. Unpublished address, American Educational Research Association, 1976. 2.16

Burstein, L. The use of data from groups for inferences about individuals in educational research. Unpublished doctoral dissertation, Stanford University, 1975. 3.11

Campbell, D. T. Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, 1958, 3, 14-25. 1.19

Cartwright, D. S. Ecological variables. In E. F. Borgatta (Ed.) *Sociological methodology, 1969*. San Francisco: Jossey-Bass, 1969. Pp. 155-218. 9.9

Cattell, R. B. An attempt at more refined definition of the cultural dimensions of syntality in modern nations. *American Sociological Review*, 1952, 17, 408-421. 9.9

Cattell, R. B. The dimensions of culture patterns by factorization of national characters. *Psychometrika*, 1949, 14, 279-298. 9.9

Cline, M. G. *et al*. Education as experimentation: Evaluation of the Follow Through Planned Variation Model. Vols. 1A, 1B. Cambridge, Mass.: Abt Associates, 1974. (ED 094890, 094891) 1.3a, 6.1, 8.3

Coleman, J. S. Methods and results in the IEA studies of effects of school on learning. *Review of Educational Research*, 45, 1975, 355-386. 2.18

Coleman, J. S., *et al*. *Equality of educational opportunity*. Washington: Government Printing Office, 1966. 1.1, 1.16, 2.18, 3.18

Cronbach, L. J. Beyond the two disciplines of scientific psychology. *American Psychologist*, 1975, 30, 116-127. 1.16, 1.21, 3.10, 10.12

Cronbach, L. J. & Gleser, G. C. *Psychological tests and personnel decisions*. Urbana: University of Illinois Press. 1957. (2nd ed., 1965) 2.8 , 10.2

Cronbach, L. J., Gleser, G. C., Nanda, ., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972. 6.2 ff.

Cronbach, L. J., Rogosa, D., Floden, R. E., & Price, G. Analysis of covariance -- Angel of salvation, or temptress and deluder? Occasional Paper, Stanford Evaluation Consortium, Stanford University, Stanford, Calif., 1976. 1.7, 1.17a, 7.2

Cronbach, L. J. & Snow, R. E. *Aptitudes and instructional methods*. New York: Irvington Publishers, 1976. Preface, 2.1 ff., 5.11, 10.2

Cronbach, L. J. & Webb, N. Between-class and within-class effects in a reported Aptitude × Treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology*, 1975, 67, 717-727. 5.1 ff.

Davis, J. A., Spaeth, J. L., & Huson, C. A technique for analysing the effects of group composition. *American Sociological Review*, 1961, 26, 215-225. 1.17a

Dogan, M. & Rokkan, S., (Eds.) Quantitative ecological analysis in the social sciences. Cambridge: M.I.T. Press, 1969. 1.10, 2.11

Duncan, O. D. Partials, partitions, and paths. In E. F. Borgatta (Ed.) Sociological methodology, 1970. San Francisco: Jossey-Bass, 1970. Pp. 38-47. 2.20

Duncan, O. D., Cuzzort, R. P., & Duncan, B. D. Statistical geography. Glencoe: Free Press, 1961. 1.10, 3.2, 9.7

Duncan, O. D., Featherman, D. L., & Duncan, B. Socioeconomic background and achievement. New York: Seminar Press, 1972. 1.10, 3.4, 3.17

Estes, W. K. The problem of inference from curves based on group data. Psychological Bulletin, 1956, 53, 134-140. 1.14

Farkas, G. Specification, residuals, and contextual effects. Sociological Methods and Research, 1974, 2, 333-364. 1.17b

Featherstone, H. J. Cognitive effects of preschool programs on different types of children. Cambridge, Mass.: Huron Institute, 1973. 2.21, 5.23 ff., 8.9 ff.

Feige, E. L., & Watts, H. W. An investigation of the consequences of partial aggregation of micro-economic data. Econometrica, 1972, 40, 343-360. 3.15

Firebaugh, G. The ecological fallacy: A reconsideration and reformulation. Unpublished manuscript, University of Indiana, 1975. 1.3, 2.12

Frank, P. (Ed.) Validation of scientific theories. New York: Collier Books, 1961. 2.24

Haney, W. The dependability of group mean scores. Unpublished paper, Harvard University Graduate School of Education, Cambridge, Mass., 1974a. 6.4

Haney, W. Units of analysis issues in the evaluation of Project Follow Through. Unpublished report, Huron Institute, Cambridge, Mass., 1974b. Preface

Hannan, M. T. Aggregation and disaggregation in sociology. Lexington, Mass.: Lexington, 1971. 1.19a, 2.11

Hannan, N., & Burstein, L. Estimation from grouped observations. American Sociological Review, 1974, 39, 374-392. 3.11

Härnqvist, K. The international study of educational achievement. In F. N. Kerlinger (Ed.) Review of Research in Education, 1975, 3, 85-109. 2.12, 9.5, 9.22

Hauser, R. M. Context and consex: A cautionary tale. *American Journal of Sociology*, 1970a, 75, 645-664. 1.17a

Hauser, R. M. Hauser replies. *American Journal of Sociology*, 1970b, 76, 517-520. 1.17 ff.

Hauser, R. M. *Socioeconomic background and educational performance*. Washington, D. C.: American Sociological Association, 1971. 1.17 ff., 2.11

Hauser, R. M. Contextual analysis revisited. *Sociological Methods and Research*, 1974, 2, 365-375. 1.17b, 6.10

Hauser, R. M., Sewell, W. H., and Alwin, D. F. High school effects on achievement. W. H. Sewell, R. M. Hauser, and D. L. Featherman (Eds.) *Schooling and achievement in American society*. New York: Academic Press, 1976. Pp. 309-341. 1.3a

Jaeger, R. M. Designing school testing programs for institutional appraisal: An application of sampling theory. Unpublished doctoral dissertation, Stanford University, 1970. 10.8

Janson, C. G. Some problems of ecological factor analysis. In M. Dogan and S. Rokkan (Eds.) *Quantitative ecological analysis in the social sciences*. Cambridge, Mass.: MIT Press, 1969. Pp. 301-342. 9.9

Kendall, P. L., & Lazarsfeld, P. F. The relation between individual and group characteristics in "The American Soldier". In P. F. Lazarsfeld & M. Rosenberg (Eds.) *The language of social research*. Glencoe, Ill.: Free Press, 1955. Pp. 290-296. 2.14

Lo, M.- Y. Statistical analysis of interaction and its application to data from the Cooperative Research Program in primary reading instruction. Unpublished doctoral dissertation, State University of New York at Buffalo. UM 73-29,111. 1973. 5.11a

Luecke, D. F., & McGinn, N. F. Regression analyses and education production functions: Can they be trusted? *Harvard Educational Review*, 1975, 45, 325-350. 2.19

Lumsden, J. Test theory. *Annual Review of Psychology*, 1976, 27, 251-280. 1.26

Maier, M. H., & Jacobs, P. I. The effects of variations in a self-instructional program on instructional outcomes. *Psychological Reports*, 1966, 18, 539-546. 2.3

Merton, R. K., & Kitt, A. Contributions to the theory of reference groups. In R. K. Merton & P. F. Lazarsfeld (Eds.) *Continuities in social structure*. Glencoe, Ill.: Free Press, 1950. 10.9

Meyer, J. W. High school effects on college intentions. *American Journal of Sociology*, 1970, 75, 59-70. 1.23, 3.18

Minkowich, A., Davis, D., & Bashi, J. An evaluation study of Israeli elementary schools. Jerusalem: Hebrew University, School of Education. To appear, 1976. 1.3a

Peckham, P. D., Glass, G. V., & Hopkins, K. D. The experimental unit in statistical analysis. *Journal of Special Education*, 1969, 3, 337-349. 1.2

Pedhazur, E. J. Analytic methods in studies of educational effects. *Review of Research in Education*, 1975, 3, 243-305. 2.20

Putnam, H. Reductionism and the nature of psychology. *Cognition*, 1973, 2, 131-146. 1.16

Ray, H. W. Final report on the Office of Economic Opportunity experiment in educational performance contracting. Unpublished report, Battelle Laboratories, Columbus, O., 1972. 1.18

Riley, M. W. *Sociological research, a case approach*. New York: Harcourt, Brace & World, 1963. 1.14, 2.12

Robinson, W. S. Ecological correlations and the behavior of individuals. *American Sociological Review*, 1950, 15, 351-357. 1.3, 2.11

Scheuch, E. K. Cross-national comparisons using aggregate data: Some substantive and methodological problems. In R. L. Merritt and S. Rokkan (Eds.) *Comparing nations: The use of quantitative data in cross-national research*. New Haven: Yale University Press, 1966. Pp. 131-167. 1.10, 1.14, 2.12

Shaycoft, M. The statistical characteristics of school means. In Flanagan, J. C., *et al*. *Studies of the American High School*. Pittsburgh: University of Pittsburgh, 1962. 6.4

Shively, W. P. "Ecological" inference: The use of aggregate data to study individuals. *American Political Science Review*, 1969, 63, 1183-1196. 1.19a

Slatin, G. T. Ecological analysis of delinquency: Aggregation effects. *American Sociological Review*, 1969, 34, 894-907. 1.3

Slatin, G. T. A factor analytic comparison of ecological and individual correlations: Some methodological implications. *Sociological Quarterly*, 1974, 15, 507-520. 9.9, 9.15

Smith, M. S. & Bissell, J. E. Report analysis: The impact of Head Start. *Harvard Educational Review*, 1970, 40, 51-104. 4.11, 5.10, 8.11 ff., 9.23

Streuning, E., & Guttentag, M. (Eds.) *Handbook of evaluation research*. Beverly Hills: Sage, 1975. Pp. 519-536. 1.24

Thorndike, E. L. On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them. American Journal of Psychology, 1939, 52, 122-124. 1.3

Tryon, R. C., & Bailey, D. E. Cluster analysis. New York: McGraw-Hill, 1970. 9.9

Welch, W., & Walberg, H. J. A national experiment in curriculum evaluation. American Educational Research Journal, 1972, 9, 373-383. 2.5

Werts, C. E. The partitioning of variance in school effects studies. American Educational Research Journal, 1968, 5, 311-318. 2.18

Westinghouse Learning Corporation. The impact of Head Start. 2 vols. Washington, D.C.: U.S. Department of Commerce, 1969. 8.11 ff.

Wittrock, M.D., & Wiley, D. E. (Eds.) The evaluation of instruction: Issues and problems. New York: Holt, Rinehart, & Winston, 1970. 1.15, 6.4

Yule, G. U., & Kendall, M. G. An introduction to the theory of statistics. London: Charles Griffin, 1950. 1.25, 2.12